



GLU - A GRADIENT MODEL FOR LANGUAGE UNIVERSALS. A COMPUTATIONAL REVISITATION OF GREENBERG UNIVERSALS

Antoni Brosa Rodríguez

ADVERTIMENT. L'accés als continguts d'aquesta tesi doctoral i la seva utilització ha de respectar els drets de la persona autora. Pot ser utilitzada per a consulta o estudi personal, així com en activitats o materials d'investigació i docència en els termes establerts a l'art. 32 del Text Refós de la Llei de Propietat Intel·lectual (RDL 1/1996). Per altres utilitzacions es requereix l'autorització prèvia i expressa de la persona autora. En qualsevol cas, en la utilització dels seus continguts caldrà indicar de forma clara el nom i cognoms de la persona autora i el títol de la tesi doctoral. No s'autoritza la seva reproducció o altres formes d'explotació efectuades amb finalitats de lucre ni la seva comunicació pública des d'un lloc aliè al servei TDX. Tampoc s'autoritza la presentació del seu contingut en una finestra o marc aliè a TDX (framing). Aquesta reserva de drets afecta tant als continguts de la tesi com als seus resums i índexs.

ADVERTENCIA. El acceso a los contenidos de esta tesis doctoral y su utilización debe respetar los derechos de la persona autora. Puede ser utilizada para consulta o estudio personal, así como en actividades o materiales de investigación y docencia en los términos establecidos en el art. 32 del Texto Refundido de la Ley de Propiedad Intelectual (RDL 1/1996). Para otros usos se requiere la autorización previa y expresa de la persona autora. En cualquier caso, en la utilización de sus contenidos se deberá indicar de forma clara el nombre y apellidos de la persona autora y el título de la tesis doctoral. No se autoriza su reproducción u otras formas de explotación efectuadas con fines lucrativos ni su comunicación pública desde un sitio ajeno al servicio TDR. Tampoco se autoriza la presentación de su contenido en una ventana o marco ajeno a TDR (framing). Esta reserva de derechos afecta tanto al contenido de la tesis como a sus resúmenes e índices.

WARNING. Access to the contents of this doctoral thesis and its use must respect the rights of the author. It can be used for reference or private study, as well as research and learning activities or materials in the terms established by the 32nd article of the Spanish Consolidated Copyright Act (RDL 1/1996). Express and previous authorization of the author is required for any other uses. In any case, when using its content, full name of the author and title of the thesis must be clearly indicated. Reproduction or other forms of for profit use or public communication from outside TDX service is not allowed. Presentation of its content in a window or frame external to TDX (framing) is not authorized either. These rights affect both the content of the thesis and its abstracts and indexes.



UNIVERSITAT ROVIRA I VIRGILI

GLU – A Gradient Model for Language Universals

A Computational Revisitation of Greenberg Universals

ANTONI BROSA RODRÍGUEZ



DOCTORAL THESIS
2024

UNIVERSITAT ROVIRA I VIRGILI

GLU - A GRADIENT MODEL FOR LANGUAGE UNIVERSALS. A COMPUTATIONAL REVISITATION OF GREENBERG UNIVERSALS

Antoni Brosa Rodríguez



UNIVERSITAT ROVIRA I VIRGILI

GLU - A GRADIENT MODEL FOR LANGUAGE UNIVERSALS

A COMPUTATIONAL REVISITATION OF GREENBERG UNIVERSALS

ANTONI BROSA RODRÍGUEZ

Supervised by

DR. MARÍA DOLORES JIMÉNEZ LÓPEZ

2024

UNIVERSITAT ROVIRA I VIRGILI

GLU - A GRADIENT MODEL FOR LANGUAGE UNIVERSALS. A COMPUTATIONAL REVISITATION OF GREENBERG UNIVERSALS

Antoni Brosa Rodríguez



UNIVERSITAT ROVIRA I VIRGILI

I STATE that the present study, entitled “GLU - A Gradient Model for Language Universals: A Computational Revisitation of Greenberg Universals”, presented by Antoni Brosa Rodríguez for the award of the degree of International Doctor, has been carried out under my supervision at the Departament de Filologies Romàniques of Universitat Rovira i Virgili and that it fulfills all the requirements for the award of International Doctor.

Tarragona, 28th of January of 2024

**María Dolores
Jiménez López**
DNI 39883298X
(TCAT)

Firmado digitalmente
por María Dolores
Jiménez López - DNI
39883298X (TCAT)
Fecha: 2024.01.28
19:49:40 +01'00'

Signature of Doctoral Thesis Supervisor

Dr. M. Dolores Jiménez López

UNIVERSITAT ROVIRA I VIRGILI

GLU - A GRADIENT MODEL FOR LANGUAGE UNIVERSALS. A COMPUTATIONAL REVISITATION OF GREENBERG UNIVERSALS

Antoni Brosa Rodríguez

Abstract

This doctoral thesis is situated in the field of Linguistic Typology and proposes an innovative approach to classify typological Universals, focusing on the presence of exceptions. Fuzzy Logic is used to understand Universality as a continuum, and a computational approach to Greenberg Universals is presented. The problem addressed concerns exceptions to Universals and the need for more clarity in relevance and gradual transition in statistical Universals.

The justification for the research is based on the revolution in Linguistic Typology introduced by Greenberg in the 1960s and the need to naturalize exceptions through terms such as “statistical Universals”. Over time, several linguists tried to improve on Greenberg’s results, but interest waned at the end of the 20th century. In the last decade, however, the rise of Computational Linguistics and access to linguistic corpora have revitalized the study of typological Universals.

The methodology adopted involves formalizing Greenberg Universals and verifying them in 143 languages using Universal Dependencies. Fuzzy Logic is employed for a stepwise classification of Universality. In addition, Grew-Match syntax and UD labels are used for formalization and structure search in linguistic corpora.

A significant theoretical contribution is presented by proposing a new classification of typological Universals using Fuzzy Logic. Our Gradient Language Universals Model (GLU) introduces a system for representing vagueness, allowing Universality to be understood as a gradual phenomenon. In this approach, any linguistic formulation within the high level is categorized as valid for describing the general behavior of languages. This innovative perspective not only improves accuracy in confirming existing Universals but also provides a solid theoretical basis for addressing linguistic variability more nuanced and comprehensively. The adoption of Fuzzy Logic represents an advance in the conceptualization of Universality, allowing for a more refined and adaptive classification of the inherent complexity of linguistic structures.

The results confirmed most of Greenberg’s Universals with greater precision, offering formulations of new Universals based on these. The proposed model’s classificatory utility is highlighted. In addition to the theoretical benefits, it is pointed out that, in the future, with more annotated linguistic data, natural language processing applications will be able to extend their findings to

currently underrepresented languages.

In summary, this thesis provides a novel perspective on the classification of typological Universals, combining Fuzzy Logic, Computational Linguistics, and the Corpus Linguistics approach to improve the accuracy and applicability of the results obtained.

Keywords – Universals in Language, Fuzzy Logic, Computational Linguistics, Greenberg, Linguistic Typology, Corpus Linguistics, Universal Dependencies, Grew-Match, Typometrics

Resumen

Esta tesis doctoral se sitúa en el ámbito de la tipología lingüística y propone un enfoque innovador para clasificar los universales tipológicos, centrándose en la presencia de excepciones. Se utiliza la lógica difusa para entender la universalidad como un continuo, y se presenta un acercamiento computacional a los universales de Greenberg. La problemática abordada se relaciona con las excepciones a los universales y la falta de claridad en términos de relevancia y transición gradual en los universales estadísticos.

La justificación de la investigación se basa en la revolución en la tipología lingüística introducida por Greenberg en los años 60 y la necesidad de naturalizar las excepciones mediante términos como “universales estadísticos”. A lo largo del tiempo, varios lingüistas intentaron mejorar los resultados de Greenberg, pero el interés disminuyó a finales del siglo XX. Sin embargo, en la última década, el auge de la lingüística computacional y el acceso a corpus lingüísticos han revitalizado el estudio de los universales tipológicos.

La metodología adoptada implica la formalización de los universales de Greenberg y su verificación en 143 lenguas utilizando Universal Dependencies. Se emplea la lógica difusa para una clasificación gradual de la universalidad. Además, se utiliza la sintaxis de Grew-Match y las etiquetas de UD para la formalización y búsqueda de estructuras en corpus lingüísticos.

Se presenta una contribución teórica significativa al proponer una nueva clasificación de los universales tipológicos mediante el uso de la lógica difusa. Nuestro Gradient Language Universals Model (GLU) introduce un sistema de representación de la vaguedad, permitiendo entender la universalidad como un fenómeno gradual. En este enfoque, cualquier formulación lingüística dentro del nivel alto es catalogada como válida para la descripción del comportamiento general de las lenguas. Esta perspectiva innovadora no solo mejora la precisión en la confirmación de los universales existentes, sino que también proporciona una base teórica sólida para abordar la variabilidad lingüística de manera más matizada y comprensiva. La adopción de la lógica difusa representa un avance en la conceptualización de la universalidad, permitiendo una clasificación más refinada y adaptable a la complejidad inherente a las estructuras lingüísticas.

Los resultados obtenidos confirman la mayoría de los universales de Greenberg con mayor

precisión. También se ofrecen formulaciones de nuevos universales basados en estos. Se destaca la utilidad clasificatoria del modelo propuesto. Además de los beneficios teóricos, se señala que, en el futuro, con más datos lingüísticos anotados, las aplicaciones de procesamiento del lenguaje natural podrán extender sus hallazgos a lenguas actualmente infrarrepresentadas.

En resumen, esta tesis aporta una perspectiva novedosa en la clasificación de universales tipológicos, combinando la lógica difusa, la lingüística computacional y el análisis de corpus para mejorar la precisión y aplicabilidad de los resultados obtenidos.

Palabras clave – Universales en el lenguaje, Lógica difusa, Lingüística computacional, Greenberg, Tipología lingüística, Lingüística de corpus, Universal Dependencies, Grew-Match, Typometrics

Resum

Aquesta tesi doctoral se situa en l'àmbit de la tipologia lingüística i proposa un enfocament innovador per a classificar els universals tipològics, centrant-se en la presència d'excepcions. S'utilitza la lògica difusa per a entendre la universalitat com un continu, i es presenta una aproximació computacional als universals de Greenberg. La problemàtica abordada es relaciona amb les excepcions als universals i la falta de claredat en termes de rellevància i transició gradual en els universals estadístics.

La justificació de la recerca es basa en la revolució en la tipologia lingüística introduïda per Greenberg en els anys 60 i la necessitat de naturalitzar les excepcions mitjançant termes com "universals estadístics". Al llarg del temps, diversos lingüistes van intentar millorar els resultats de Greenberg, però l'interès va disminuir a finals del segle XX. No obstant això, en l'última dècada, l'auge de la lingüística computacional i l'accés a corpus lingüístics han revitalitzat l'estudi dels universals tipològics.

La metodologia adoptada implica la formalització dels universals de Greenberg i la seva verificació en 143 llengües utilitzant Universal Dependencies. S'empra la lògica difusa per a una classificació gradual de la universalitat. A més, s'utilitza la sintaxi de Grew-Match i les etiquetes de UD per a la formalització i cerca d'estructures en corpus lingüístics.

Es presenta una contribució teòrica significativa amb la proposta de una nova classificació dels universals tipològics mitjançant l'ús de la lògica difusa. El nostre Gradient Language Universals Model (GLU) introdueix un sistema de representació de la vaguetat, permetent entendre la universalitat com un fenomen gradual. En aquest enfocament, qualsevol formulació lingüística dins del nivell alt és catalogada com a vàlida per a la descripció del comportament general de les llengües. Aquesta perspectiva innovadora no només millora la precisió en la confirmació dels universals existents, sinó que també proporciona una base teòrica sòlida per a abordar la variabilitat lingüística de manera més matisada i comprensiva. L'adopció de la lògica difusa representa un avenç en la conceptualització de la universalitat, permetent una classificació més refinada i adaptable a la complexitat inherent a les estructures lingüístiques.

Els resultats obtinguts confirmen la majoria dels universals de Greenberg amb major precisió.

També s’ofereixen formulacions de nous universals basats en aquests. Es destaca la utilitat classificatòria del model proposat. A més dels beneficis teòrics, s’assenyala que, en el futur, amb més dades lingüístiques anotades, les aplicacions de processament del llenguatge natural podran estendre les seves troballes a llengües actualment infrarepresentades.

En resum, aquesta tesi aporta una perspectiva nova en la classificació d’universals tipològics, combinant la lògica difusa, la lingüística computacional i l’anàlisi de corpus per a millorar la precisió i aplicabilitat dels resultats obtinguts.

Paraules clau – Universals en el llenguatge, Lògica difusa, Lingüística computacional, Greenberg, Tipologia lingüística, Lingüística de corpus, Universal Dependencies, Grew-Match, Typometrics

Acknowledgments

First, I would like to thank Dr. M. Dolores Jiménez López for everything she has done for me. Her dynamic management style has allowed me to reach my final goal most appropriately and fruitfully. Her demands, wisdom, and willingness to work have allowed me to grow as a researcher and reach places unimaginable a few years ago. Without her, the thesis would have been very different (or would not exist). For all these reasons, I consider her a mentor, and she has made this thesis a starting point, not an endpoint.

I must thank, in general, the members of the Departament de Filologies Romàniques of the Universitat Rovira i Virgili for their kindness in welcoming me and their trust in me. This should also be extended to the Doctoral Programme in Humanistic Studies. I also thank the financial support derived from the Martí i Franquès contract (2019PMF-PIPF-99).

The GRLMC (Research Group in Mathematical Linguistics) has always been there for everything I have needed and has listened patiently to my concerns. I must especially thank my colleagues Luca Clementi and Susana Campillo with gratitude.

Special mention should be made of another of its members, Dr. Adrià Torrens, who has also followed the course of this doctoral thesis with particular interest. His advice has positively impacted me and has been valuable and effective. He has been a kind and generous pioneer.

I must thank Dr. Nicola Grandi of the Università di Bologna for accepting me for a research stay in his Department. He has allowed me to go much deeper into Linguistic Typology, solve existing doubts with the trichotomous model, and raise new ones. Besides him, I have been able to exchange feelings with very friendly typologists such as Dr. Simone Mattioli and to get to know closely the PhD students of this Department, whose admiration is both academic and personal. Thanks to Dr. Sebastiano Moruzzi for lending me several books on fuzzy logic that have been useful. I am also grateful to Dr. Harald Hammarström for giving me access to his tools.

It is equally important to recognize that this thesis would be different without the generosity of Dr. Sylvain Kahane. He has always been available to resolve linguistic, formal, or computational doubts. In addition, he has challenged me as a researcher and made me go beyond my original

goals. During my research stay in France, I have felt like a member of the Université Paris Nanterre, and it has allowed me to discuss my progress with leading figures in the field, such as Dr. Kim Gerdes and Dr. Natalia Levshina, who have found time for me. I must not forget the doctoral students of the MoDyCo research group and their kindness. I must strongly highlight the help of Santiago Herrera and his patience and altruism in solving my doubts about Grew-Match. Also, thanks to Eskil Blaafat Mundal for his willingness to talk about Norwegian and its conditionals.

Thanks to the various colleagues who, during my Erasmus as a lecturer at the University of Rouen and the University of Lodz, have listened to me. Also, to Dr. Laura Becker, who has offered her opinion in several conversations about my research at a crucial moment of my thesis. I would also like to extend my gratitude to all the attendees with constructive criticism at various conferences and journal reviewers who have contributed to improving the research details or confidence in what is being done.

Personally, I must thank my family for their support, especially my mother, Dr. M. José Rodríguez Campillo. She is ultimately responsible for creating the ideal environment for me to pursue a venture like this one. It is a great fortune always to have her infinite kindness. Thanks also to my brother David for putting up with me. Thanks to my friends for allowing me to disconnect from this research (if possible). Thanks to Pati for her unconditional support and for sharing many happy moments.

I dedicate this thesis to my grandmother and grandfather.

Publications Related to this PhD Dissertation

2021

1. Brosa-Rodríguez, A. (2021). Corpus y Universales lingüísticos en pragmática. *E-Aesla (Centro Virtual del Instituto Cervantes)*, 7, 37-49.
2. Torrens-Urrutia, A., Jiménez-López, M. D. & Brosa-Rodríguez, A. (2021). A Fuzzy Approach To Language Universals for NLP. *IEEE CIS International Conference on Fuzzy Systems 2021*, 992-997. Metrics: Web of Science CPCI-S / Core:A

2022

3. Brosa-Rodríguez, A. & Jiménez-López, M. D. (2022). La jerarquía implicacional de Berlin y Kay a partir de frecuencias. *E-Aesla (Centro Virtual del Instituto Cervantes)*, 8, 3-19.
4. Torrens-Urrutia, A., Jiménez-López, M. D., Brosa-Rodríguez, A. & Adamczyk, A. (2022). A Fuzzy Grammar for Evaluating Universality and Complexity. *Mathematics*, 10(15), 1-23. Metrics: Impact Factor 2.592, Citescore 2.9, Scopus Q1, JCR Q1. Featured Paper Distinction.
5. Ramírez-Pérez, R. & Brosa-Rodríguez, A. (2022). An Intercultural Study of Universals in Pragmatics: From Polish to Spanish. *Wschód-Zachód: Literatura i Język na styku kultur*, Akademia Pomorska w Słupsku, 131-142.

2023

6. Brosa-Rodríguez, A. & Jiménez-López, M. D. (2023). A Typometrical Study of Greenberg's Linguistic Universal 1. *Lecture Notes in Networks and Systems*, 741, Springer Nature, 186-196. Metrics: Indexed in DBLP; INSPEC, SCOPUS, SCImago, Norwegian Register for Scientific Journals and Series, WTI Frankfurt eG, zbMATH.

7. Brosa-Rodríguez, A. (2023). Principales dificultades en la investigación de Universales lingüísticos. *Recerca en Humanitats 2022*, Publicacions URV, 23-38.

2024

8. Brosa-Rodríguez, A. & Jiménez-López, M. D. (2024). Quantifying Basic Colours' Salience from Cross-linguistic Corpora. *Colors Research & Application*, 49(1), 34-50. Metrics: Impact Factor 1.4, Citescore 3.3, Scopus Q3, JCR Q3. Distinguished to be included in the Special Collection Color Cognition.

9. Brosa-Rodríguez, A., Jiménez-López, M. D. & Torrens-Urrutia A. (2024). Exploring the Complexity of Natural Languages: A Fuzzy Evaluative Perspective on Greenberg Universals. *AIMS Mathematics*, 9(1), 2181-2214. Metrics: Impact factor 2.2, Citescore 3, Scopus Q1, JCR Q1.

Accepted & In Press

10. Brosa-Rodríguez, A. & Kahane, S. (2024). New Proposal of Greenberg's Universal 14 from Typometrics. LREC-Coling. The Association for Computational Linguistics. Metrics: Web of Science CPCI-S / Core:A/B

11. Brosa-Rodríguez, A. (2024). Distintos métodos digitales para revisar Universales lingüísticos. Accepted & In Press. Metrics: SPI (Q1).

12. Brosa-Rodríguez, A. & Rodríguez-Campillo, M. J. (2024). A computational typological characterization of a Turkish-German Code-Switching, Springer Nature. Accepted & In Press. Metrics: SPI (Q1).

13. Brosa-Rodríguez, A. (2024). Universales en las lenguas: sus distintas definiciones y una nueva propuesta gradual. *Recerca en Humanitats 2023*, Publicacions URV.

14. Brosa-Rodríguez, A. (2024). Análisis de herramientas computacionales para trabajar en tipología lingüística con Universal Dependencies. *Recerca en Humanitats 2024*, Publicacions URV.

Contents

List of Figures	xix
List of Tables	xxiii
I Introduction	1
1 Motivation	3
2 Hypothesis, Research Questions, and Objectives	13
3 Methodology	15
4 Thesis Outline	21
II State of the Art	27
5 Definition of Universals of Language	29
5.1 Universals According to Hockett	30
5.2 Universals According to Greenberg	31
5.3 Universals According to Chomsky	32
5.4 Other Definitions of Universals	33
5.5 Summary of the Definition of Universals in Language	37
6 Terminology Used in Universals of Language	39
6.1 Labels to Distinguish the Conception of Universal	40
6.2 Labels to Distinguish the Structure of the Universal	47
6.3 Labels to Distinguish the Presence of Exceptions	51
6.4 Labels that Are Used for Different Variables	61
7 Methodologies for Universals Research	65
7.1 Universals Formulated from the Armchair	66
7.2 Universals Formulated from the Laboratory	73
7.3 Universals Formulated from the Field	75
8 Universals & Linguistic Level	79
8.1 Syntactic Universals	80
8.2 Morphological Universals	81
8.3 Semantic Universals	82
8.4 Pragmatic Universals	83
8.5 Phonological Universals	84
9 Universals in The Universals Archive	85
9.1 Classification According to the Terminology Distinction	86
9.2 Classification According to the Methodology	88
9.3 Classification According to the Linguistic Level	89

10	Historical Panorama of the Universals	91
10.1	Universals concerning Greenberg	92
10.1.1	Universals before Greenberg (1963)	92
10.1.2	Universals in Greenberg (1963)	97
10.1.3	Universals after Greenberg (1963)	102
10.2	Computational Universals of Language	111
10.3	Typology and Universals	124
11	Limitations in Typological Universals Research	137
11.1	Unawareness of Total Languages of the World	138
11.2	Dialects and Languages Distinction	139
11.3	Dead Languages and Creoles	140
11.4	Scarcity of Data	142
11.5	Anglocentrism	144
11.6	Sampling	148
III	Theoretical Framework	155
12	Typological Universals Frame	157
12.1	Data Type	159
12.2	Greenberg Universals	161
13	Token-Based Typology	167
13.1	Universal Dependencies Project	168
13.2	Benefits of Universal Dependencies	171
13.3	Limitations of Universal Dependencies	172
13.4	Tools to Analyze Data from Universal Dependencies	174
13.4.1	Scalable and Efficient Tree Search	175
13.4.2	Prague Markup Language - Tree Query	176
13.4.3	Kontext	179
13.4.4	Graph Rewriting for NLP - Match	181
13.4.5	Infrastructure for the Exploration of Syntax and Semantics	185
13.4.6	Choice	187
13.4.7	Typometrics	188
14	Fuzzy Logic	195
14.1	Gradience and Typology	196
14.2	Gradience and Universals	197
14.3	The Sorites Paradox	198
14.4	Fuzzy Natural Logic Concepts	199
IV	A Gradient Language Universals Model - GLU	203
15	Fuzzy Logic Models	205
15.1	Five Triangular Fuzzy Sets	205
15.2	Five Trapezoidal Fuzzy Sets	207
15.3	From Five to Three Fuzzy Sets	208
16	Fuzzy Logic Model for Language Universals	211
16.1	Description of the Model	212
16.2	Example of Application to the Universal 13	214
16.3	Comparison of Traditional Labels and the GLU Model	217
16.4	Benefits of the GLU Model	219
16.5	On Possible Objections to the GLU Model	222

V	A Computational Approach to Greenberg’s Universals	225
17	Formal Prerequisites	227
17.1	Universal Dependencies Tags	228
17.2	Surface Universal Dependencies Syntactic Relations	241
17.3	Grew-Match Formalisation Syntax	243
17.4	Data Management Formalisations	244
18	Decisions in Formalisation	245
18.1	Basic Word Order	245
18.2	Adpositions	246
18.3	Subjects, Objects, and Verbs	250
18.4	Adjectives	252
19	Data Management	253
19.1	Sample	253
19.2	Data and Limitations	255
19.3	Script	258
19.4	Data Cleaning	258
19.5	Calculating Universal 1	260
19.6	Implicational Formulae	265
20	Universals Classification According to the Tool	269
20.1	Universals through Typometrics	271
20.2	Universals through Grew-Match	272
20.3	Universals through UD Treebank Hub Page	273
20.4	Universals through Combination of Different Methods	274
20.5	Universals with Current Limitations	274
21	Formal Analysis of Universals	277
21.1	Universal 1	279
21.1.1	How to Check It in Typometrics	279
21.1.2	Formalisation Proposed for Grew-Match	280
21.1.3	Graphs from Typometrics	292
21.1.4	Trichotomous Expression Based on Grew-Match Data	296
21.2	Universal 2	299
21.2.1	How to Partially Check the Lemma Clustering	299
21.2.2	Formalization Proposed for Grew-Match	301
21.2.3	Trichotomous Expression Based on Grew-Match Data	305
21.3	Universal 3	307
21.3.1	How to Check It in Typometrics	307
21.3.2	Formalization Proposed for Grew-Match	309
21.3.3	Graphs from Typometrics	314
21.3.4	Trichotomous Expression Based on Grew-Match Data	318
21.4	Universal 4	320
21.4.1	How to Check It in Typometrics	320
21.4.2	Formalization Proposed for Grew-Match	321
21.4.3	Graphs from Typometrics	324
21.4.4	Trichotomous Expression Based on Grew-Match Data	327
21.5	Universal 5	330
21.5.1	How to Partially Check it in Grew-Match	330
21.5.2	Trichotomous Expression Based on Grew-Match Data	333
21.6	Universal 6	335
21.6.1	Formalization Proposed for Grew-Match	335
21.6.2	Trichotomous Expression Based on Grew-Match Data	337
21.7	Universal 7	339
21.7.1	How to partially check it in Typometrics	339

21.7.2	Formalization Proposed for Grew-Match	340
21.7.3	Trichotomous Expression Based on Grew-Match Data	344
21.8	Universal 8	346
21.8.1	Current Limitation: Data no Available	346
21.9	Universal 9	346
21.9.1	Current Limitation: Data no Available	346
21.10	Universal 10	347
21.10.1	Current Limitation: Data no Available	347
21.11	Universal 11	348
21.11.1	Formalization Proposed for Grew-Match	348
21.11.2	Trichotomous Expression Based on Grew-Match Data	355
21.12	Universal 12	357
21.12.1	Formalization Proposed for Grew-Match	357
21.12.2	Limitations Based on Grew-Match Data	359
21.13	Universal 13	360
21.13.1	How to Check It in Typometrics	360
21.13.2	Formalization Proposed for Grew-Match	360
21.13.3	Graphs from Typometrics	367
21.13.4	Trichotomous Expression Based on Grew-Match Data	369
21.14	Universal 14	371
21.14.1	Checking Combining PanLex and Grew-Match	371
21.14.2	Trichotomous Expression Based on Grew-Match Data and Typometrics . .	376
21.15	Universal 15	382
21.15.1	Current Limitation: Data no Available	382
21.16	Universal 16	383
21.16.1	How to Check It in Typometrics	383
21.16.2	Formalization Proposed for Grew-Match	384
21.16.3	Graphs from Typometrics	389
21.16.4	Trichotomous Expression Based on Grew-Match Data	392
21.17	Universal 17	394
21.17.1	How to Check It in Typometrics	394
21.17.2	Formalization Proposed for Grew-Match	395
21.17.3	Graphs from Typometrics	401
21.17.4	Trichotomous Expression Based on Grew-Match Data	406
21.18	Universal 18	408
21.18.1	Formalization Proposed for Grew-Match	408
21.18.2	Trichotomous Expression Based on Grew-Match Data	415
21.19	Universal 19	417
21.19.1	How to Check It in Typometrics	417
21.19.2	Formalization Proposed for Grew-Match	418
21.19.3	Graphs from Typometrics	421
21.19.4	Trichotomous Expression Based on Grew-Match Data	427
21.20	Universal 20	429
21.20.1	Formalization Proposed for Grew-Match	429
21.20.2	Trichotomous Expression Based on Grew-Match Data	441
21.21	Universal 21	443
21.21.1	How to Check It in Typometrics	443
21.21.2	Formalization Proposed for Grew-Match	443
21.21.3	Graphs from Typometrics	449
21.21.4	Trichotomous Expression Based on Grew-Match Data	452
21.22	Universal 22	455
21.22.1	Formalization Proposed for Grew-Match	455
21.22.2	Limitations Based on Grew-Match Data	461
21.23	Universal 23	462
21.23.1	How to Partially Check it in Grew-Match	462
21.23.2	Trichotomous Expression Based on Grew-Match Data	467

21.24	Universal 24	469
21.24.1	How to Check It in Typometrics	469
21.24.2	Formalization Proposed for Grew-Match	471
21.24.3	Graphs from Typometrics	477
21.24.4	Trichotomous Expression Based on Grew-Match Data	480
21.25	Universal 25	483
21.25.1	How to Check It in Typometrics	483
21.25.2	Formalization Proposed for Grew-Match	484
21.25.3	Graphs from Typometrics	486
21.25.4	Trichotomous Expression Based on Grew-Match Data	489
21.26	Universal 26	491
21.26.1	Current Limitation: Data no Available	491
21.27	Universal 27	491
21.27.1	Current Limitation: Data no Available	491
21.28	Universal 28	492
21.28.1	Current Limitation: Data no Available	492
21.29	Universal 29	492
21.29.1	Current Limitation: Data no Available	492
21.30	Universal 30	492
21.30.1	Extracting the Information in the Treebank Hub Page	492
21.30.2	Formalization Proposed for Grew-Match	493
21.30.3	Trichotomous Expression Based on Grew-Match Data	498
21.31	Universal 31	499
21.31.1	Formalization Proposed for Grew-Match	499
21.31.2	Trichotomous Expression Based on Grew-Match Data	507
21.32	Universal 32	509
21.32.1	Formalization Proposed for Grew-Match	509
21.32.2	Trichotomous Expression Based on Grew-Match Data	519
21.33	Universal 33	521
21.33.1	Formalization Proposed for Grew-Match	521
21.33.2	Limitations Based on Grew-Match Data	529
21.34	Universal 34	530
21.34.1	Extracting the Information in the Treebank Hub Page	530
21.34.2	Formalization Proposed for Grew-Match	530
21.34.3	Trichotomous Expression Based on Grew-Match Data	532
21.35	Universal 35	533
21.35.1	Current Limitation: Data no Available	533
21.36	Universal 36	533
21.36.1	Extracting the Information in the Treebank Hub Page	533
21.36.2	Formalization Proposed for Grew-Match	534
21.36.3	Trichotomous Expression Based on Grew-Match Data	535
21.37	Universal 37	536
21.37.1	Formalization Proposed for Grew-Match	536
21.37.2	Trichotomous Expression Based on Grew-Match Data	538
21.38	Universal 38	539
21.38.1	Current Limitation: Data no Available	539
21.39	Universal 39	540
21.39.1	Current Limitation: Data no Available	540
21.40	Universal 40	540
21.40.1	Formalization Proposed for Grew-Match	540
21.40.2	Trichotomous Expression Based on Grew-Match Data	549
21.41	Universal 41	551
21.41.1	How to Check it Generating Tables in Grew-Match	551
21.41.2	Formalization Proposed for Grew-Match	552
21.41.3	Trichotomous Expression Based on Grew-Match Data	558
21.42	Universal 42	559

21.42.1	Extracting the Information in the Treebank Hub Page	559
21.42.2	Formalization Proposed for Grew-Match	559
21.42.3	Trichotomous Expression Based on Grew-Match Data	562
21.43	Universal 43	563
21.43.1	Extracting the Information in the Treebank Hub Page	563
21.43.2	Formalization Proposed for Grew-Match	563
21.43.3	Trichotomous Expression Based on Grew-Match Data	565
21.44	Universal 44	567
21.44.1	Formalization Proposed for Grew-Match	567
21.44.2	Trichotomous Expression Based on Grew-Match Data	572
21.45	Universal 45	574
21.45.1	Formalization Proposed for Grew-Match	574
21.45.2	Trichotomous Expression Based on Grew-Match Data	575
22	General Results	577
22.1	Quantitative Review of All Greenberg Universals	577
22.2	Quantitative Review of the Proposed High-Universals	578
VI	Conclusions	581
23	Conclusions	583
23.1	GLU: A Gradient Concept of Universality	584
23.1.1	Review and Compilation of the Universals Terminology	585
23.1.2	A Theoretical Proposal: GLU Model	587
23.2	Formalization of Greenberg Universals	589
23.2.1	Advantages and Limitations for a Computational Study of Greenberg Uni- versals	589
23.2.2	A Formalization of Greenberg's Universals	594
23.3	Analysis of Greenberg Universals	595
23.3.1	Testing Universals Computationally with Real Texts	595
23.3.2	Formulation of New Universals in Language	597
24	Future Work	601
24.1	Future Research on Universality as a Gradient Concept	601
24.2	New Possibilities on the Formalization of Greenberg Universals	602
24.3	Future Work on the Analysis of Greenberg Universals	603
	Appendix	605
	Bibliography	607

List of Figures

1.1	Yearly Evolution of Papers with “Universals in Language” in the Title in the Web of Science	4
1.2	Yearly Evolution of Papers with “Linguistic Universals” in the Title in the Web of Science	4
1.3	Yearly Evolution of Papers with “Linguistic Typology” in the Title in the Web of Science	5
1.4	Yearly Evolution of Papers within Computer Science with “Linguistic Universals” in the Title in the Web of Science	5
1.5	Distribution by Field of Papers with “Linguistic Typology” in the Title in the Web of Science	6
1.6	Distribution by Language of Papers with “Universals in Language” in the Title in the Web of Science	7
3.1	Steps in the Research	16
4.1	Outline of the Parts of the Thesis	25
5.1	Comparison between a Conception of Universal as the Possible and Universal as the Probable in Dryer (1998).	37
6.1	Word Cloud of the Terminology We Have Found	40
6.2	Implicational Hierarchy from Berlin & Kay (1969)	50
9.1	Statistical and Absolute Universals Distribution on The Universals Archive . .	88
9.2	Implicational and Non-implicational Universals Distribution on The Universals Archive	88
9.3	Universals Distribution on The Universals Archive According to the Linguistic Level	90
13.1	Universal Dependencies Website	169
13.2	Treebank Annotation Example in Raw	170
13.3	General View of “SETS Treebank Search” Tool	176
13.4	General View of “PML Tree Query” Tool	177
13.5	Query in “PML Tree Query” Tool	178
13.6	General View of “Kontext” Tool	180
13.7	General View of “Grew-Match Match” Tool	182
13.8	Example of the Table on <i>obj</i> Dependency’s Part of Speech in Spanish (AnCora) .	183
13.9	Example of Table on Case Types & Frequencies in Languages/Treebanks . . .	184
13.10	Example of Adpositions in an English Treebank, Clustered by Frequency of Lemma	184
13.11	Example of Occurrences of Adjectives before and after the Noun Head in Catalan	185
13.12	General View of “INESS” Tool	186
13.13	General View of Typometrics	190
13.14	Example of Results in Typometrics Capturing the Relation between Nominal Objects and Adpositions	192
15.1	Model with 5 Triangular Fuzzy Sets in Gajovic, Kerkez, and Kocovic (2018) . .	206

15.2	Model with 5 Trapezoidal Fuzzy Sets in Moreno Cárdenas, Zapata-Zapata, and Kim (2020)	207
15.3	Trichotomous Expression Proposed by Torrens-Urrutia, Novák, and Jiménez-López (2022)	209
16.1	Gradient Language Universals Model	212
16.2	Trichotomous Expression of the GLU-M for Universal 13	215
19.1	Example of the Checking of Significance of our Data	263
21.1	Nominal Subjects and Objects Direction Checking	293
21.2	Nominal Subjects and Objects Distribution Proposal	294
21.3	Pronominal Subjects and Objects Distribution Checking	295
21.4	Pronominal Subjects and Objects Distribution Proposal	295
21.5	Trichotomous Expression for Universal 1	297
21.6	Trichotomous Expression for Universal 2	305
21.7	Nominal Subjects and Adpositions with Nouns Direction Checking	314
21.8	Nominal Subjects and Adpositions with Nouns Direction Proposal	315
21.9	Nominal Subjects and Adpositions with Pronouns Direction Checking	316
21.10	Nominal Subjects and Adpositions with Pronouns Direction Proposal	316
21.11	Nominal Subjects and Adpositions with Verbs Direction Checking	317
21.12	Nominal Subjects and Adpositions with Verbs Direction Proposal	317
21.13	Trichotomous Expression for Universal 3	318
21.14	Nominal Objects and Adpositions with Nouns Direction Checking	324
21.15	Nominal Objects and Adpositions with Nouns Direction Proposal	325
21.16	Nominal Objects and Adpositions with Pronouns Direction Checking	326
21.17	Nominal Objects and Adpositions with Pronouns Direction Proposal	327
21.18	Nominal Objects and Adpositions with Verbs Direction Checking	328
21.19	Trichotomous Expression for Universal 4	328
21.20	Trichotomous Expression for Universal 5	334
21.21	Trichotomous Expression for Universal 6	338
21.22	Trichotomous Expression for Universal 7	344
21.23	Trichotomous Expression for Universal 11	356
21.24	Distance of O and VD to V Head in Spanish (AnCora)	366
21.25	Distance of O and VD to V Head in Classical Chinese (Kyoto)	366
21.26	Nominal Direct Objects and Verbal Direct Objects Direction Checking	368
21.27	Nominal Direct Objects and Verbal Direct Objects Direction Proposal	368
21.28	Trichotomous Expression for Universal 13	370
21.29	Percentage of IF-Clauses before Conclusions	379
21.30	Dispersion of Percentage of VO Structures and IF-before Structures	379
21.31	Dispersion of Percentage of IF-Clauses before and ADV-Clauses before	380
21.32	Trichotomous Expression for Universal 14	380
21.33	Subjects and Auxiliaries Direction Checking	389
21.34	Subjects and Auxiliaries Direction Proposal	390
21.35	Objects and Auxiliaries Direction Checking	391
21.36	Trichotomous Expression for Universal 16	392
21.37	Subjects and Adjectives (Nominal Modifiers) Direction Checking	402
21.38	Subjects and Adjectives (Nominal Modifiers) Direction Proposal	403
21.39	Adjectives and Adverbs (Adjectival Modifiers) Direction Checking	404
21.40	Adjectives and Adverbs (Adjectival Modifiers) Direction Proposal	404
21.41	Nominal Objects and Adverbs (Adjectival Modifiers) Checking	405
21.42	Nominal Objects and Adverbs (Adjectival Modifiers) Proposal	406
21.43	Trichotomous Expression for Universal 17	407
21.44	Trichotomous Expression for Universal 18	416
21.45	Adjectives Modifying Nouns Direction Checking	422
21.46	Adjectives Modifying Nouns Direction Proposal	423

21.47	Adjectives Modifying Pronouns Direction Checking	425
21.48	Adjectives Modifying Proper Nouns Direction Checking	426
21.49	Trichotomous Expression for Universal 19	428
21.50	Trichotomous Expression for Universal 20	442
21.51	Adverbs Modifying Adjectives and Adjectives Modifying Nouns Direction Checking	450
21.52	Adverbs Modifying Adjectives and Adjectives Modifying Nouns Direction Proposal	451
21.53	Adverbs Modifying Adjectives and Direct Objects Direction Checking	451
21.54	Adverbs Modifying Adjectives and Direct Objects Direction Proposal	452
21.55	Trichotomous Expression for Universal 21	453
21.56	Trichotomous Expression for Universal 23	467
21.57	Verbs Modifying Nouns and Adposition Direction Checking	477
21.58	Verbs Modifying Nouns and Adposition Direction Proposal	478
21.59	Verbs Modifying Nouns and Adjectives Modifying Nouns Direction Checking .	479
21.60	Verbs Modifying Nouns and Adjectives Modifying Nouns Direction Proposal .	480
21.61	Trichotomous Expression for Universal 24	481
21.62	Nominal (Common) and Pronominal Direct Objects Direction Checking	487
21.63	Nominal (Common) and Pronominal Direct Objects Direction Proposal	487
21.64	Nominal (Proper) and Pronominal Direct Objects Direction Checking	488
21.65	Nominal (Proper) and Pronominal Direct Objects Direction Proposal	489
21.66	Trichotomous Expression for Universal 25	490
21.67	Trichotomous Expression for Universal 30	498
21.68	Trichotomous Expression for Universal 31	508
21.69	Trichotomous Expression for Universal 32	520
21.70	Trichotomous Expression for Universal 34	532
21.71	Trichotomous Expression for Universal 36	535
21.72	Trichotomous Expression for Universal 37	539
21.73	Trichotomous Expression for Universal 40	550
21.74	Trichotomous Expression for Universal 41	558
21.75	Trichotomous Expression for Universal 42	562
21.76	Trichotomous Expression for Universal 43	566
21.77	Trichotomous Expression for Universal 44	573
21.78	Trichotomous Expression for Universal 45	575
22.1	Trichotomous Expression for Greenberg's Original Universals	578
22.2	Proposal of the Trichotomous Expression for Greenberg's Universals	579
23.1	Example of a Trichotomous Expression for Language Universals	589

UNIVERSITAT ROVIRA I VIRGILI

GLU - A GRADIENT MODEL FOR LANGUAGE UNIVERSALS. A COMPUTATIONAL REVISITATION OF GREENBERG UNIVERSALS

Antoni Brosa Rodríguez

List of Tables

5.1	Differences in the Different Views of Universals in Language	38
6.1	Summary of Labels Regarding the Conception of Universals	64
6.2	Summary of Labels Regarding the Structure of Universals	64
6.3	Summary of Labels Regarding the Exceptions of Universals	64
13.1	Summary of the Characteristics of the Analyzed Tools	193
17.1	Universal Part of Speech	228
17.2	Features	229
17.3	Pronominal Type	230
17.4	Nominal Type	230
17.5	Possessive	230
17.6	Reflexive	231
17.7	Foreign Word	231
17.8	Abbreviation	231
17.9	Typo or Misspelled Word	231
17.10	Gender	232
17.11	Animacy	232
17.12	Noun Class	233
17.13	Number	234
17.14	Case	235
17.15	Definiteness or State	236
17.16	Degree	236
17.17	Form of Verb or Deverbative	237
17.18	Mood	237
17.19	Tense	238
17.20	Aspect	238
17.21	Voice	239
17.22	Evidentiality	239
17.23	Polarity	239
17.24	Person	240
17.25	Politeness	240
17.26	Clusivity	240
17.27	Own Dependencies	241
17.28	Complement Specific Dependencies	241
17.29	Shared Dependencies with Universal Dependencies	242
17.30	SUD Deep Features	242
17.31	Grew-Match Formalisation Syntax	243
17.32	Data Management Formalisations	244
19.1	Percentage of Languages Order in the Sample	254
19.2	Comparison between 1000 Sentences Limit or without Limit	257
20.1	Summary of Tested Universals, their Value and Modifications	270
20.2	Overview of Universals and Tools	276

21.1	Technical Decisions for Checking Universal 1 in Typometrics	279
21.2	Accomplishment of Different Formalizations of Universal 1 in a Quantitative- Typometrical Way	298
21.3	Summary of Characteristics of the Universal 1 Analysis	298
21.4	Summary of Characteristics of the Universal 2 Analysis	306
21.5	Technical Decisions for Checking Universal 3 in Typometrics	307
21.6	Summary of Characteristics of the Universal 3 Analysis	319
21.7	Technical Decisions for Checking Universal 4 in Typometrics	320
21.8	Summary of Characteristics of the Universal 4 Analysis	329
21.9	Summary of Characteristics of the Universal 5 Analysis	335
21.10	Summary of Characteristics of the Universal 6 Analysis	338
21.11	Technical Decisions for Checking Universal 7 in Typometrics	340
21.12	Summary of Characteristics of the Universal 7 Analysis	345
21.13	Summary of Characteristics of the Universal 11 Analysis	356
21.14	Technical Decisions for Checking Universal 13 in Typometrics	360
21.15	Summary of Characteristics of the Universal 13 Analysis	370
21.16	Summary of Characteristics of the Universal 14 Analysis	381
21.17	Technical Decisions for Checking Universal 16 in Typometrics	383
21.18	Summary of Characteristics of the Universal 16 Analysis	393
21.19	Technical Decisions for Checking Universal 17 in Typometrics	394
21.20	Summary of Characteristics of the Universal 17 Analysis	408
21.21	Summary of Characteristics of the Universal 18 Analysis	417
21.22	Technical Decisions for Checking Universal 19 in Typometrics	417
21.23	Summary of Characteristics of the Universal 19 Analysis	428
21.24	Summary of Characteristics of the Universal 20 Analysis	442
21.25	Technical Decisions for Checking Universal 21 in Typometrics	443
21.26	Summary of Characteristics of the Universal 21 Analysis	454
21.27	Summary of Characteristics of the Universal 23 Analysis	468
21.28	Technical Decisions for Checking Universal 24 in Typometrics	469
21.29	Summary of Characteristics of the Universal 24 Analysis	482
21.30	Technical Decisions for Checking Universal 25 in Typometrics	483
21.31	Summary of Characteristics of the Universal 25 Analysis	490
21.32	Summary of Characteristics of the Universal 30 Analysis	499
21.33	Summary of Characteristics of the Universal 31 Analysis	508
21.34	Summary of Characteristics of the Universal 32 Analysis	521
21.35	Summary of Characteristics of the Universal 34 Analysis	533
21.36	Summary of Characteristics of the Universal 36 Analysis	536
21.37	Summary of Characteristics of the Universal 37 Analysis	539
21.38	Summary of Characteristics of the Universal 40 Analysis	551
21.39	Summary of Characteristics of the Universal 41 Analysis	559
21.40	Summary of Characteristics of the Universal 42 Analysis	563
21.41	Summary of Characteristics of the Universal 43 Analysis	567
21.42	Summary of Characteristics of the Universal 44 Analysis	573
21.43	Summary of Characteristics of the Universal 45 Analysis	575

Part I

Introduction

UNIVERSITAT ROVIRA I VIRGILI

GLU - A GRADIENT MODEL FOR LANGUAGE UNIVERSALS. A COMPUTATIONAL REVISITATION OF GREENBERG UNIVERSALS

Antoni Brosa Rodríguez

1

Motivation

In the vast and intricate realm of language, a fundamental question has intrigued linguists and philosophers throughout the centuries: the existence of linguistic Universals. These Universals represent recurring patterns, structures, and concepts that transcend cultural and geographical diversities, raising fascinating questions about the inherent nature of human language. In this context, our study focuses on exploring the topic of language Universals, checking and unraveling the common threads that connect diverse languages of the world. Despite advances in linguistic research, exploring linguistic Universals remains an evolving field. This study aims to contribute to the existing literature by analyzing linguistic Universals and offering new perspectives and approaches that enrich our understanding of the topic.

An understandable question that may arise concerns the contemporary relevance of studying typological Universals. Since its inception, this field has been accompanied by debates about its reliability, utility, and existence, which persist. It is important to note that the concept of Universals is used in different disciplines and even different schools within linguistics. This disparity can lead to confusion about what we mean when discussing Universals. This study focuses on the capacity to categorize languages into types, offering valuable insights into their behavior rather than a simplistic notion of universality. The significance of this approach has been convincingly demonstrated, as we will explore in detail within this thesis.

From a quantitative perspective, it is possible to see the importance of the subject of this thesis. For example, data from the Web of Science (which we take as a reference point) can be used. The total number of publications about the topic cannot be obtained, as there are a large number of tags that refer to it, such as “Universals in language” (16,338 papers in this resource containing this

expression in the title), “Linguistic Universals” (4,264) or “Linguistic Typology” (3,140), among others. However, we believe it is interesting to see how, despite being a relatively standard topic, its scientific interest does not diminish, as can be seen in the figures 1.1, 1.2 and 1.3. Note that 2023 is not considered as it is the year in which this data is captured and, therefore, shows inconclusive data that generates bias. However, one can see the trend culminating in 2022. The choice to restrict the search to the titles of certain expressions causes the number of occurrences of papers related to the topic to be lower than it is.

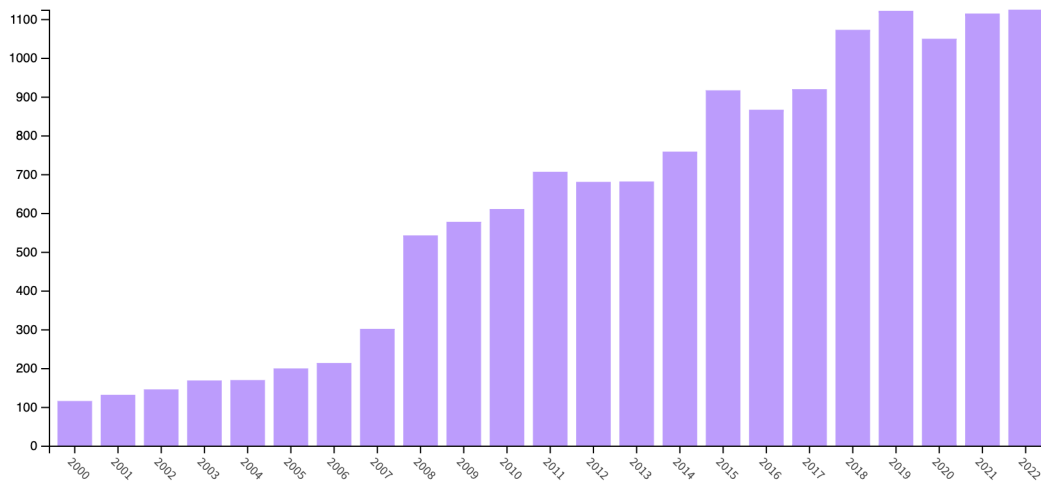


Figure 1.1: Yearly Evolution of Papers with “Universals in Language” in the Title in the Web of Science

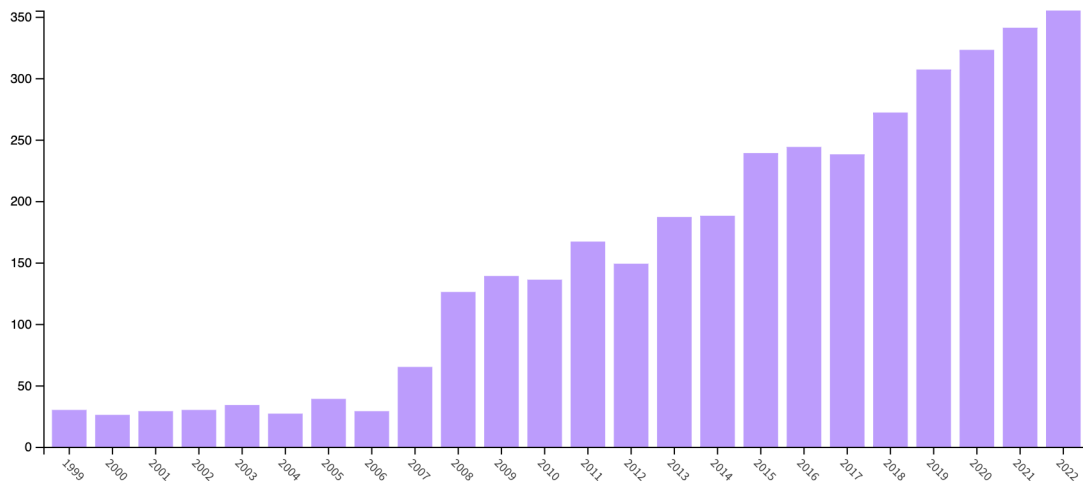


Figure 1.2: Yearly Evolution of Papers with “Linguistic Universals” in the Title in the Web of Science

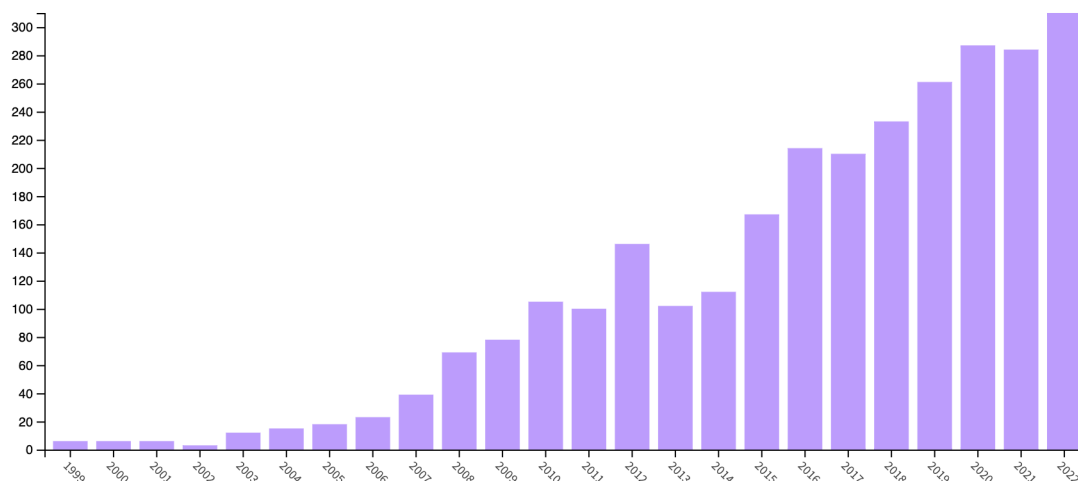


Figure 1.3: Yearly Evolution of Papers with “Linguistic Typology” in the Title in the Web of Science

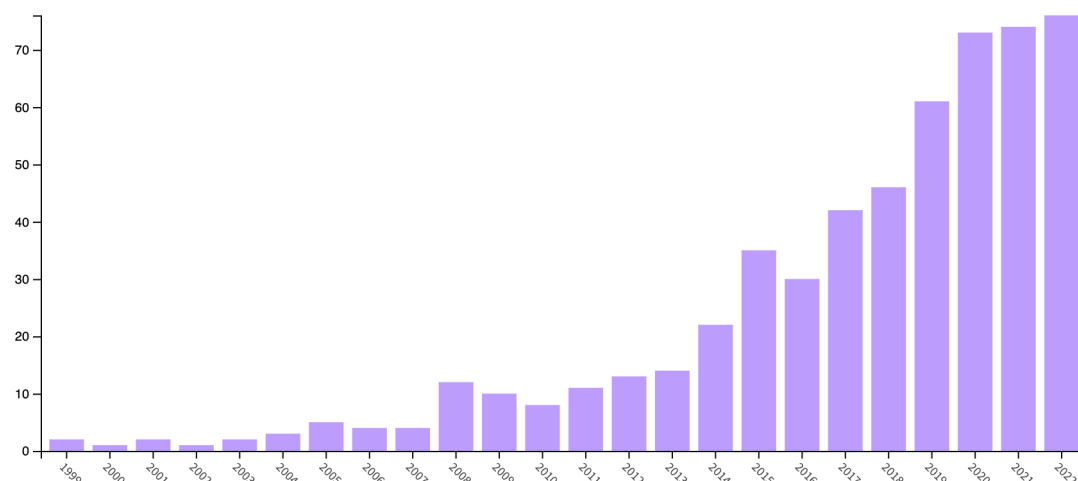


Figure 1.4: Yearly Evolution of Papers within Computer Science with “Linguistic Universals” in the Title in the Web of Science

Another possibility is that the Web of Science allows us to obtain more precise results. By way of example, in figure 1.4, we wanted to check whether this growing number of publications containing the subject of our research also occurs in computer science. This field is also crucial for this thesis. As can be seen, the same trend can be observed in the four graphs we show in figures 1.1, 1.2, 1.3 and 1.4: interest is growing (at least quantitatively). Observing this typical pattern can help us conclude that interest in the subject has not waned, even though we cannot obtain total and exact numbers due to the terminological disparity in the field. Furthermore, we believe showing the trend or paradigm shift towards 2009 is relevant. The quantitative growth of papers is remarkable and coincides temporally with the emergence of greater interdisciplinarity between Linguistic Typology and Computational Linguistics.

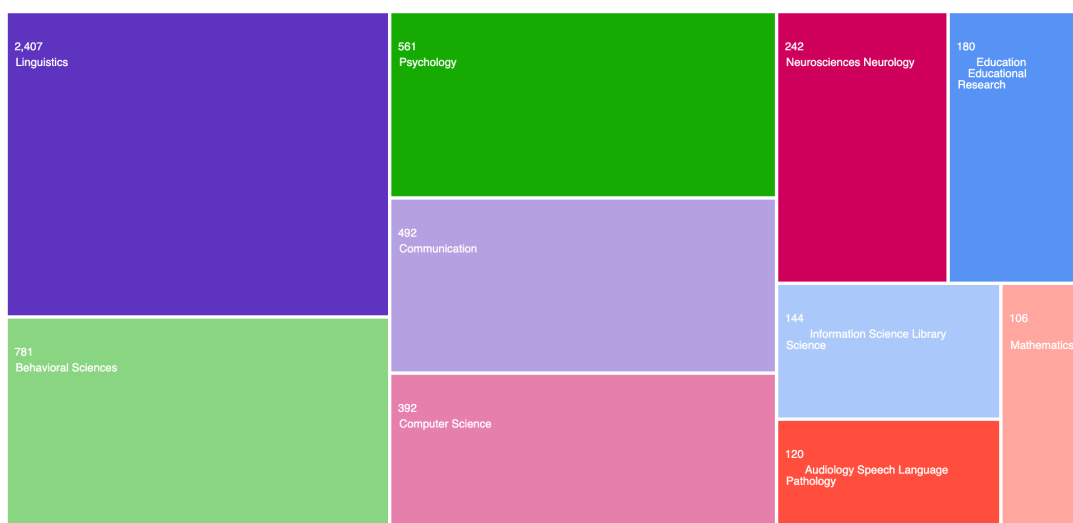


Figure 1.5: Distribution by Field of Papers with “Linguistic Typology” in the Title in the Web of Science

We also want to check the areas to which the study of Universals in language is related. This information may point to the subject’s most common points of interdisciplinarity. As an example, since all the results are similar, we look at the behavior of the papers with “Linguistic Typology” in figure 1.5.

Most of the papers are linguistic. However, the topic’s importance concerning other fields, such as those shown in the figure, particularly in the behavioral sciences and psychology, must be considered. There is a clear interest in the relationship between Universals and the cognitive or psychological causes behind them.

Finally, we also want to show the distribution of papers by language. Again, since the distributions are very similar in any of the labels consulted, we show by example the language of the papers with “Universals in language” in the title in figure 1.6. The dominant language in Universals in language studies is English. The following language (at a great distance) is Russian. Therefore, English is assumed to be the usual language of scientific communication.

The intrinsic value of understanding linguistic Universals goes beyond the confines of academic inquiry and quantitative terms. It has become a powerful tool with far-reaching implications for language education. Identifying common patterns in languages enables educators to tailor teaching approaches that resonate with these shared structures, thereby enhancing the efficacy of language instruction. This streamlines the learning process and fosters a deeper appreciation for the universality that binds diverse linguistic expressions.

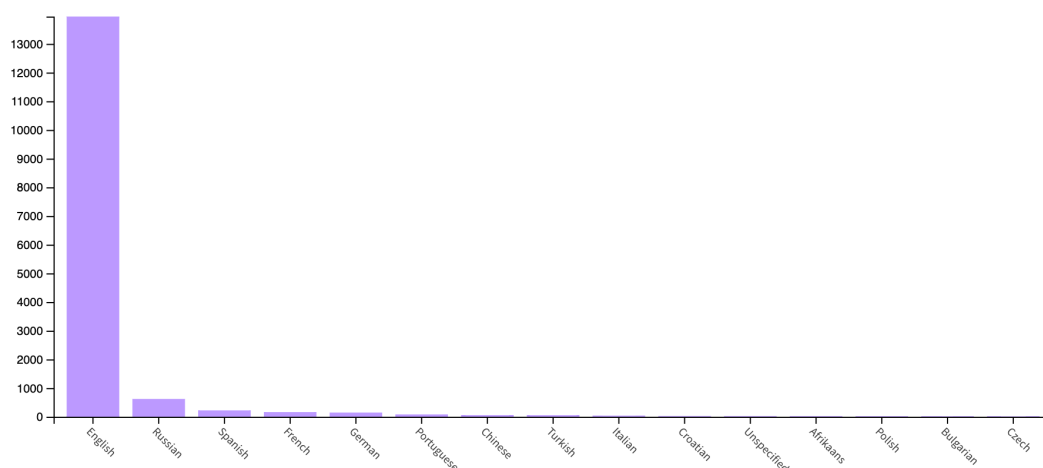


Figure 1.6: Distribution by Language of Papers with “Universals in Language” in the Title in the Web of Science

Practically speaking, the insights derived from the study of Universals significantly contribute to developing language-related technologies. One notable example is the advancement of automatic translation systems. Understanding the commonalities in language structures enables the creation of more sophisticated algorithms that can accurately translate content across various languages. This has transformative implications for global communication, breaking language barriers and fostering a more interconnected world.

In essence, the study of Universals transcends the realm of theoretical linguistics; it emerges as a dynamic field with tangible implications for education, communication, and technology (Xia, Han, Palmer, & Joshi, 2001). Embracing this exploration not only enriches our understanding of the intricacies of language but also empowers us to harness its Universals aspects for practical, real-world applications.

In the age of Natural Language Processing (NLP) transforming linguistics, Universals become crucial. NLP, revolutionizing how we deal with language, faces a fundamental challenge: while languages share basic features at an abstract level, world language structures show much variation. This diversity makes it challenging to transfer NLP models across languages or create widely applicable systems. As a result, NLP tech is primarily available in a few well-resourced languages, leaving others underserved. Understanding linguistic variety is essential to developing practical multilingual NLP applications and making NLP tech more globally accessible. Linguistic Typology aims to identify invariances or order within the variation in the structures of different languages worldwide, as previously mentioned. The best summary of the main challenge can be found in the

quote from Ponti et al. (2019:1):

Languages may share Universals features at a deep, abstract level, but the structures found in real-world, surface-level natural language vary significantly. This variation makes it challenging to transfer NLP models across languages or to develop systems that apply to a wide range of languages. Consequently, the availability of NLP technology is limited to a handful of resource-rich languages, leaving many other languages behind. Understanding linguistic variation systematically is crucial for developing practical multilingual NLP applications, thus making NLP technology more accessible globally.

We are not referring to using information derived from Universals for extensively studied languages like English. Unfortunately, many advancements mentioned before are only accessible for a select number of major European languages. Speakers of languages that have received limited attention in scholarly research lack the application of these tools, thus missing out on potential benefits. For instance, in the context of a parser (Tomita & Carbonell, 1987), if we can ascertain that a less-studied language shares syntactic similarities with English at a rate of 98%, we can leverage an English parser to comprehend and extract information from it. This information can be advantageous for advancing machine translation and other applications. The insights into shared linguistic structures among languages, as dictated by linguistic Universals, offer valuable knowledge. For example, if we know that a language follows a basic SVO order, it implies a likelihood of having prepositions and various other features associated with this linguistic type. Consequently, minimal information about a language can yield substantial knowledge.

Authors such as Naseem, Chen, Barzilay, and Johnson (2010) have shown that using the knowledge generated by Universals allows better results in dependency parsing. Thanks to this type of resource, in Universal Dependencies (Nivre et al., 2023), we can obtain exciting data for formulating Universals. This mutual implication invites us to affirm that we are dealing with a symbiosis. Advances in Universals make it possible to improve NLP (in terms of extension to more languages or refinement of models). However, NLP makes it possible to enhance the Universals that can be obtained or to obtain new ones (Merlo, 2015).

Also, in the concrete aspect of annotation, Universals allow for several improvements that make the process more efficient. The best example is the thesis by Garrette (2015). It argues for a middle way of annotation, i.e., automatic supervised annotation. The Universals allow that with few indications from an annotator/supervisor; much data can be annotated, which he/she will subsequently only have to supervise. This is clearly assumed by Garrette (2015:9):

To obtain the best weakly-supervised NLP tools we can, it is essential that we make the best possible use of all language information we have available; this means both efficiently collecting the right kinds of annotations from people and combining them with the right Universals inductive biases.

The transfer of information from a well-studied language to a poorly resourced one through Typology acting as a bridge is richly captured by O’Horan, Berzak, Vulić, Reichart, and Korhonen (2016:1297):

The field of Linguistic Typology offers valuable resources for nearing both of these theoretical ideals: it studies and classifies the world’s languages according to their structural and functional features, with the aim of explaining both the common properties and the structural diversity of languages. Many of the current popular solutions to multilingual NLP: transfer of information from resource-rich to resource-poor languages (...), joint multilingual learning (...), and development of Universals models (...), either assume or explicitly make use of information related to Linguistic Typology.

Different tools use typological knowledge to benefit low-resourced languages. Both Garrette (2015) and O’Horan et al. (2016) also point to a clear economic benefit. For all these languages that are not the most implemented, they are not without data either; we have some information, but the various computational tools still need to be applied. This is mainly due to economic and market aspects. Correct annotation needs at least human supervision. The time spent on corrections will be less if this process is carried out satisfactorily. Universals can help in this corpus tagging task, which reduces effort and costs. In the case of languages without rich data, this process goes beyond the economic aspect, as it provides us with information that has not been available until now (O’Horan et al., 2016:1302):

the main area of NLP in which information from structural Typology has been exploited thus far is multilingual dependency parsing. In this task, a priori information about the predominant orderings of syntactic categories across languages are used to guide models when parsing a resource-poor language and using training data from other languages. This information is available in typological resources (e.g., WALS) which, among a variety of other syntactic features, list the dominant word orderings for many languages

While in Futrell, Levy, and Gibson (2020) computational typological data are used to generate this knowledge applicable or extensible to more languages, O’Horan et al. (2016) highlight the use of classical typological data. In other words, it is shown how information from grammar-based databases allows the grouping of languages by type (Universals implications) and, therefore, facilitates the determination of the similarity between languages for the prediction of the behavior of a language (O’Horan et al., 2016:1297):

Another line of work seeks to increase the coverage of typological information using existing information in typological databases. Daumé III and Campbell (2007) and Bakker (2008) use existing WALS features to learn typological implications of the kind pioneered by Greenberg (1963). Such rules can be used to predict unknown feature values for new languages. Georgi, Xia, and Lewis (2010) use documented WALS features to cluster languages, and subsequently predict new feature values using nearest-neighbour projection. A classifier-based approach for predicting new feature values from documented WALS information is presented in Takamura, Nagata, and Kawasaki (2016). Coke, King, and Radev (2016) predict word order typological features by combining documented typological and genealogical features with the multilingual alignment approach discussed above

Classical typological information is generally more extensive (more languages are documented). A more accurate representation of linguistic phenomena compensates for the smaller information size in computational Typology. As indicated by Ponti et al. (2019:3): “This survey explores the idea of a new approach to Typology that guarantees gradient in its categories through continuous representations, and bottom-up development of typological information from raw data through automatic inference”. In other words, more precise symmetries can be drawn between languages by similarity, thanks to greater detail in their characterization. The labels used in grammars can sometimes be too simple and fail to capture correctly the richness within a given type (more rigid or more flexible, for example).

This information contained in Universals can also be used for other objects of study in linguistics. It is possible, for example, to relate Universals to linguistic complexity (Bentz, Gutierrez-Vasques, Sozinova, & Samardžić, 2022; Coloma, 2017; Mi Oh & Pellegrino, 2022). Having this typological information available from quantitative data from occurrences of different corpora allows us to obtain exciting metrics impacting complexity studies. The clearest example of this between the (inversely proportional) relationship between these two objects of study is found in Torrens-Urrutia, Jiménez-López, Brosa-Rodríguez, and Adamczyk (2022).

In addition, there are industrial applications of the knowledge generated by Universals. Jameson and Webster (2019:1037), for example, demonstrate how Berlin and Kay’s implicational hierarchy benefits a much more appropriate technological development:

Impacts of this paradigm are also found in areas of industrial standards, environmental design applications, and display device design. By increasing knowledge of how color similarity and semantics are interwoven, we are more aware of how we might enlist color preference variation across groups to personalize color display settings to improve user experience with recent advances in display and user-interface technologies. These are issues that bear on international standards for lighting and environmental design and preference, as well as individual variation in color perception.

Another evident application of the advantages derived from the study of Universals is found in second language teaching (Ozeki & Shirai, 2007). From an absolute standpoint, information is accessible regarding what is universally applicable and what exists exclusively in specific languages, as evident in implicational hierarchies. This enables the reorganization of curricula with these considerations in mind, aiming to facilitate learning to the greatest extent possible. For instance, learners can leverage the knowledge of universally observed linguistic patterns to build a foundation before delving into language-specific nuances (Eckman, 2010; Eckman, Bell, & Nelson, 1988). From a relative perspective, adapting foreign language teaching classes based on the learner’s mother tongue and target language becomes feasible, thanks to the typological knowledge at our disposal.

This allows for a more tailored approach, considering linguistic differences between the learner's native and target languages, thus optimizing the language learning process.

It is also important to note that the link between Language Universals and Second Language Learning can also be understood as a certain kind of symbiosis (Greenberg, 1991; Ramat, 2009). In addition to all the positive aspects observed in applying knowledge of Universals in Second Languages, there are benefits in the opposite direction. The productions obtained in second language speakers are very interesting samples to see the behavior of Universals in such productions. In other words, one can observe the actual incidence of all the laws promoting the emergence of these Universals.

Regarding mother tongue acquisition, some authors also see the usefulness of Universals as apparent (Ferguson, 1978b). In particular, Hurford (2009:49) points out that "Given a Universals: 'If a language has X, it always has Y,' then X cannot be acquired unless Y has been acquired first as a way of stating the relation between child language and Universals". In this paper, therefore, the parallelism is drawn on the understanding that "individual languages evolve along lines very similar to the development of language in children" (Hurford, 2009:48). Therefore, starting from a Universals such as "there is no language with a trial number that does not have a dual number", it can be inferred that children cannot learn trial if they have not first learned dual. However, according to Bowerman (2010), this issue has yet to be developed (concerning Typology).

In short, the study of Universals significantly contributes in several respects. Its initial aim of allowing the classification of languages from a linguistic perspective has led to a better understanding of the functioning of human languages, both from a more linguistic point of view and to the cognitive implications that this entails. Recently, attempts have been made to take advantage of the information generated in studies on Universals in fields such as language acquisition, second language learning, or computation.

Precisely, Computational Linguistics offers new tools and perspectives that allow the object of study to be approached through new methodologies. This field is in constant progress, and much must be done. Both linguists and experts in the more computational part can significantly contribute to the field and improve our knowledge of the phenomenon.

UNIVERSITAT ROVIRA I VIRGILI

GLU - A GRADIENT MODEL FOR LANGUAGE UNIVERSALS. A COMPUTATIONAL REVISITATION OF GREENBERG UNIVERSALS

Antoni Brosa Rodríguez

2

Hypothesis, Research Questions, and Objectives

The study we have carried out offers a new perspective on Greenberg Universals from a twofold methodological approach. Our hypothesis is the following: the integration of fuzzy logic in the classification of Greenbergian Universals and the utilization of digitized corpora for revisiting these Universals based on authentic text data is anticipated to yield a natural and richer representation of these in a more accessible way.

This proposal has arisen after a critical reading of the state of the art, where different limitations have been detected in different aspects. The following questions have emerged after the analysis:

1. Where is the cut-off point between something universal with exceptions and something that is not universal?
2. How can the presence or absence of proven exceptions to the Universals be formulated to be more fairly represented?
3. Is it possible to test or validate Greenberg Universals computationally?
4. Can we formulate new Universals from Greenberg ones?

In order to be able to test the hypothesis and answer the questions we have formulated, we have decided to set the following objectives:

1. To review the concept of Universals in Language.
2. To propose a new classification of Universals in Language, with special attention to the presence of exceptions.
3. To provide a formalization for Greenberg Universals.
4. To computationally check Greenberg Universals through real texts.
5. To revisit the formulation of some Greenberg Universals.

3

Methodology

This doctoral thesis adopts a straightforward interdisciplinary approach. In it, aspects of Linguistic Typology, fuzzy logic, and Computational Linguistics are combined to offer a possible new solution to the problem observed with the exceptions, to know in a more detailed and formal way the Greenberg Universals, and to offer other possible Universals. The advantage of an interdisciplinary proposal is the novel solution to problems of one discipline thanks to another. This type of synergy allows scientific progress. However, it is also necessary to make explicit that this thesis is not a work on fuzzy logic or Computational Linguistics. These are not the object of study; we take advantage of models or tools from these disciplines to offer possible solutions to the problems pointed out. This is quite clear when reviewing the state of the art of the object of study (Greenberg Universals). In other words, we will not know these disciplines in depth in this thesis. However, they will appear as fundamental aspects in the methodological development to obtain the objectives set.

Likewise, a classical typologist may find certain limitations in the study presented. The proposed approach does not guarantee perfect representativeness of the world's languages or other representativeness issues. The researcher cannot solve this as it is an inherent limitation of the approach. This aspect could be solved in the mid-term with a greater availability of languages in Universal Dependencies, which could qualify (or not) the findings in this research. This thesis does not attempt to compete with the results obtained from grammars but rather to offer an interpretation of the same phenomenon from another approach or methodology since there is room for both (Haddington, Rostila, & Tuomarla, 2005). This allows a greater linguistic knowledge, multilevel, and a series of advantages that make up for the shortcomings of the different methodologies.

The verification of the presence of the phenomena captured in the Universals through occurrences in corpora makes this work with a quantitative methodology. We reserve the pure quantitative result and its distribution for the claims represented in Typometrics as an initiation in a research path that we believe will gain strength in the future, without this being the primary objective of this thesis. However, since we have set out to contrast Greenberg’s Universals through this corpus methodology, converting the quantitative data into linguistic labels is necessary, which will favor the comparability between our results and those of Greenberg.

In summary, the steps we have carried out in the study to arrive at the characterization in a trichotomous expression of Greenberg’s Universals are represented in figure 3.1.

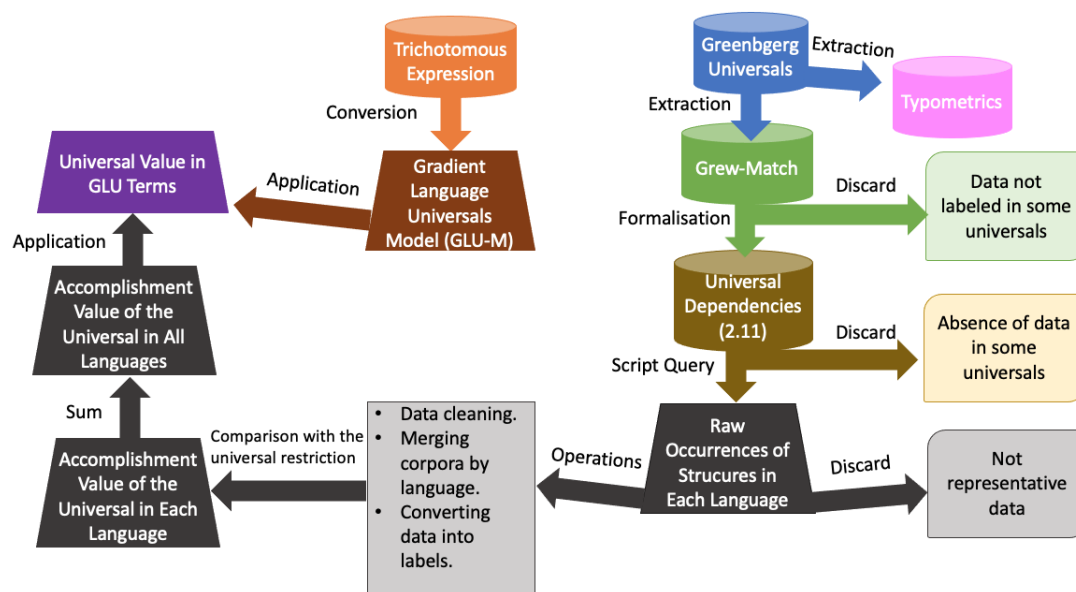


Figure 3.1: Steps in the Research

We attach importance to the method we present, which allows us a (different) approach to Greenberg’s Universals. We believe in the potential of an approach based on real texts of different languages. We understand that it is a valid and interesting way of doing linguistics through what is known as typometrics or token-based Typology. Our literature review has not identified any approach in computation that revisits Greenberg’s Universals holistically. Therefore, we emphasize our review of the Universals and the computational limitations we show in reviewing them in their entirety, something that also does not exist to our knowledge. The linguistic description between natural language and the formalization we propose of these Universals (some of them never formalized before) is also attractive, which is often forgotten in more computational studies and impacts the results, as we will show. Thanks to the improvement of data availability, in the

future, it will be possible to add more languages to those analyzed, using the same methodology that we offer, and to assess whether there is any Universal we have validated or proposed that is no longer a Universal.

1. About fuzzy logic, an existing classification model is adopted and adapted: trichotomous expression. It opposes the Boolean perspective and understands the phenomena analyzed and classified as a continuum. We rely on a model provided by Novák (2008a); Novák, Perfilieva, and Dvorak (2016): the Trichotomous Expression. This is proposed to model linguistic expressions and vagueness, as occurs with the sorites paradox. In this thesis, this model is adapted to the phenomenon under study: universality. This material applied to universality can be seen in Torrens-Urrutia, Jiménez-López, and Brosa-Rodríguez (2021). The model we created from the trichotomous expression is the Gradient Language Universals Model. Once we have our model, we move on to the concrete study of Greenberg Universals. However, since they are independent processes, this process could have been taken last, although this is not the case.
2. Concerning the revisit of Universals, the whole procedure has been done computationally. We access Greenberg's original publication and extract the Universals he formulated, a list of 45 Universals (Greenberg, 1963).
3. Once we have these, we review, one by one, which ones can be testable using the new Typometrics (Gerdes, Kahane, & Chen, 2021) tool. The number of testable Universals is low, and their testability needs to be improved. Therefore, it is not a tool that allows, at present, a reasonable translation of Greenberg's Universals to the methodology we propose in this thesis. However, given the relevance of the graphs it generates and the possibilities it opens up in terms of a purely quantitative reading of the linguistic phenomena being analyzed, together with its simplicity, we maintain the graphs we can create.
4. Subsequently, we convert the natural language (English) with which the Greenberg Universals have been formulated into a formal language. In addition to disambiguating some English expressions contained in the Universals, it allows the use of the Grew-Match tool (Guillaume, 2021) for their further formalization. Therefore, we rely on the syntax and tags that exist in Grew-Match. The syntax is the tool's own syntax. The tags are those of the Universal Dependencies annotation scheme (Nivre et al., 2023) and Surface Universal Dependencies (Gerdes, Guillaume, Kahane, & Perrier, 2022; Kahane, Vanhove, Ziane, & Guillaume, 2021). In this step, we discard some Universals to which formalization cannot be proposed with the

mentioned scheme. Such information has yet to be tagged in Universal Dependencies (UD). Therefore, no data relating to these Universals can be obtained.

5. Once all the Universals have been formalized, a document that unifies them (a script) is elaborated. The script is run in Grew-Match, and we load all the corpora of different languages in Universal Dependencies (in the most updated version at the time of research, 2.11). This version has over 250 different corpora, corresponding to approximately 150 languages. These languages are very varied, with a tendency towards Eurasian languages. These corpora are of different types and sizes. The analysis of the corpora as a whole is not technically feasible due to the large number of operations to be carried out. We opted for an analysis of the first 1000 sentences of each corpus, which guarantees replicability. Analyzing 1000 sentences for each corpus means that almost 250 million different operations are carried out in this thesis, hence the limitation. A comparison of the possible bias between the selection of 1000 sentences and the whole corpus is also conducted, and no differences are observed.
6. Once the approximately 250 UD corpora (2.11) have been analyzed, we discard some Universals for which no data have been obtained. This is due to the high specificity of the Universals, which causes us not to observe occurrences of some complex structures in real texts. This gives us an output with the different occurrences of the different formalizations referring to the structures contained in the Greenberg Universals. After applying a representativeness test ($p < 0.05$) to the data, we discard the corpora in which the result cannot be certified as reliable separately in each Universals.
7. Subsequently, we carry out further operations to process the results obtained. We clean the data. We merge the corpora referring to the same language. We convert the occurrences obtained into labels to guarantee comparability with the Greenberg claim in natural language. To carry out this process, we follow the proposals of Gries (2013) and Levshina et al. (2021).
8. Once the above operations have been carried out, we must obtain the value of compliance with the Greenberg Universals in each language. To do this, the structure obtained in each language is compared with the structure contained in the Greenberg claim. Once the value of compliance or non-compliance with the Universals in each of the languages is available, the results can be added up to obtain a total value in relation to the universe of languages analyzed.
9. Since we have the number of compliance with the Universals in our set, we can now use this result for its representation in the Trichotomous Expression from our Gradient Language

Universals Model. The Universal is placed at the corresponding point of the scale. It is considered a valid claim if the Universal appears in the High-Universals group.

UNIVERSITAT ROVIRA I VIRGILI

GLU - A GRADIENT MODEL FOR LANGUAGE UNIVERSALS. A COMPUTATIONAL REVISITATION OF GREENBERG UNIVERSALS

Antoni Brosa Rodríguez

4

Thesis Outline

This doctoral thesis is structured in six main parts.

1. **Part I:** The first part is the one we are finishing with this chapter, the *Introduction*. It is divided into:
 - **Chapter 1. Motivation:** It shows the need to study Universals in language today.
 - **Chapter 2. Hypothesis, Research Questions, and Objectives:** It sets out the research hypothesis, the questions we have formulated, and our objectives.
 - **Chapter 3. Methodology:** It explains the research methodology followed and the key steps.
 - **Chapter 4. Thesis Outline:** This chapter shows an orderly summary of the parts of the thesis.
2. **Part II:** The second part, *State of the Art*, presents a literature review related to the study topic. It includes the necessary information that supports and helps to locate and understand the research being carried out. It is divided into:
 - **Chapter 5. Definition of Universals of Language:** The different definitions of Universals are analysed.
 - **Chapter 6. Terminology Used in Universals of Language:** We offer an analysis of the different labels associated with the concept of Universals (about its conception from the definition taken, from the structure in which it is formulated, and in relation to the presence or absence of exceptions).

- **Chapter 7. Methodologies for Universals Research:** We show the different methodologies that can be used when studying Universals in language and their characteristics.
 - **Chapter 8. Universals & Linguistic Level:** We offer a review of Universals regarding the linguistic level.
 - **Chapter 9. Universals in The Universals Archive:** We quantitatively overview how many Universals distinguished by the classifications in the previous chapters are in The Universals Archive.
 - **Chapter 10. Historical Panorama of the Universals:** We offer a historical journey into the study of Universals. The figure of Greenberg has been chosen as the pivotal figure. We also show a summary of the analysis of Universals in the last decade utilizing corpora. It closes with an overview of the evolution of the relationship between Typology and Universals.
 - **Chapter 11. Limitations in Typological Universals Research:** This section summarizes the different limitations involved in the study of Typological Universals and aspects that influence their research.
3. **Part III:** The third part, *Theoretical Framework*, provides the theoretical basis for the research development. It is divided into:
- **Chapter 12. Typological Universals Frame:** Firstly, the theoretical basis of the research concerning Linguistic Typology is selected and set out. In other words, we highlight the definition of Universal followed, the terminology involved, the methodology followed, the linguistic level involved, the Universals analyzed, and the limitations we assume to be inherent to the study we are carrying out.
 - **Chapter 13. Token-Based Typology:** Secondly, we provide essential information on Computational Linguistics and, specifically, on the Universal Dependencies tool.
 - **Chapter 14. Fuzzy Logic:** Thirdly, we make explicit the notions of fuzzy logic involved in our proposal and its general conception.
4. **Part IV:** The fourth part, *A Gradient Language Universals Model - GLU*, gathers the model we propose in this doctoral thesis. It is divided into:
- **Chapter 15. Fuzzy Logic Models:** It shows and comments on the existing basic possibilities in fuzzy logic that can be used to build our model.

-
- **Chapter 16. Fuzzy Logic Model for Language Universals:** It explains all the information about the fuzzy logic contained in it and how the Gradient Language Universals Model works. An example of practice with the analysis of a Universals is shown. The use of this model is compared with the labels identified in the literature. The benefits of applying the Gradient Language Universals Model are reviewed. Some possible doubts about the limitations of the model are resolved.
5. **Part V:** The fifth part, *A Computational Approach to Greenberg Universals*, contains the development and results of the analysis of Greenberg Universals in Universal Dependencies. We assess which Universals are technically possible to analyze in UD and which are not. Next, we formalize these Universals. Then, we design a study (with different methodologies) using corpus data in the available languages. We offer the proposal for the formalization of Greenberg's Universals and explain how this is done. It is divided into:
- **Chapter 17. Formal Prerequisites:** It contains the terminology used in the formalisms to be carried out. The annotation system and the syntax used are shown.
 - **Chapter 18. Decisions in Formalisation:** It shows the fundamental aspects of our proposed formalization methodology.
 - **Chapter 19. Data Management:** It explains the different operations carried out to manage and process the raw data obtained.
 - **Chapter 20. Universals Classification According to the Tool:** We provide a classification of Greenberg Universals in relation to the different methodologies that can be used to analyze them with Universal Dependencies data. We also show which Universals cannot currently be tested and why. There is a systematic review and classification of all Greenberg Universals that also has a comprehensive and computational approach. There are many approaches to some particular Greenberg Universals. However, we are unaware of complete and deep reviews of the Greenberg Universals understood as a whole. We offer a differentiation between which Universals are possible to analyze with the availability of current data and which are not. This novel classification lets us know where we stand and allows us to focus on improving those aspects of Greenberg's proposal that we cannot yet reach computationally.
 - **Chapter 21. Formal Analysis of Universals:** It shows, in detail, the analysis of each of the 45 Universals. This includes, first, the formalization we propose to capture their occurrences. All the patterns we will present are new creations that we propose as formalizations of Greenberg Universals (and other proposals that go further). The

interest of the formalizations we offer is twofold. On the one hand, it allows us to obtain in formal language Universals that may be ambiguous in some terms in natural language. Therefore, we are presenting mathematically the Universals proposed by Greenberg. On the other hand, this formalization is necessary to be part of the plot that will be used to search for Universals through Grew-Match and to be able to extract quantitative results from the UD corpora. We offer a guide to check these Universals classified through different practical methods or tools for a linguist. Although Grew-Match can check all Universals available in UD, we want to show, where possible, the use of other tools that are useful for a typologist for different reasons (e.g., less computational knowledge requirement or better visualization of the data). Subsequently, the results obtained and the trichotomous expression corresponding to these results are shown as the main findings.

- **Chapter 22. General Results:** It is devoted to a general quantitative review of the results obtained. We show the data relating to the revision of Greenberg's Universals, the Universals that can be posited from Greenberg's Universals, and the implications that this has.

6. **Part VI:** The sixth part, *Conclusions*, concludes and reviews the research regarding the objectives. It is divided into:

- **Chapter 23. Conclusions:** It contains the actual conclusions organized around the different contributions of this thesis.
- **Chapter 24. Future Work:** It contains possible avenues for future research from this doctoral thesis.

We have divided the thesis into parts corresponding to different stages or actions. This allows for an independent reading of any of the parts if desired. This order has been proposed based on the dependence of the parts and their research process. A summary of this can be seen in figure 4.1.

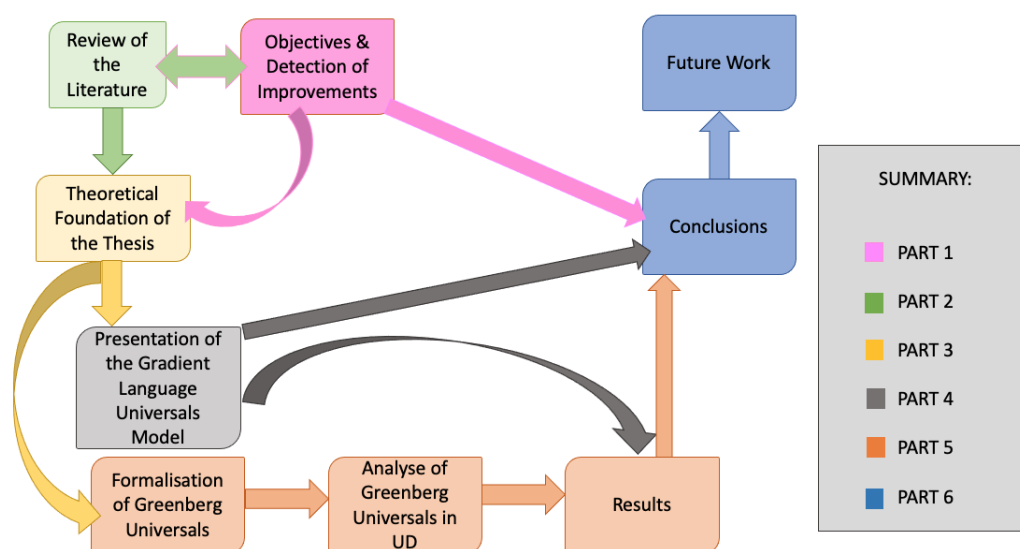


Figure 4.1: Outline of the Parts of the Thesis

UNIVERSITAT ROVIRA I VIRGILI

GLU - A GRADIENT MODEL FOR LANGUAGE UNIVERSALS. A COMPUTATIONAL REVISITATION OF GREENBERG UNIVERSALS

Antoni Brosa Rodríguez

Part II

State of the Art

UNIVERSITAT ROVIRA I VIRGILI

GLU - A GRADIENT MODEL FOR LANGUAGE UNIVERSALS. A COMPUTATIONAL REVISITATION OF GREENBERG UNIVERSALS

Antoni Brosa Rodríguez

5

Definition of Universals of Language

Contents

5.1	Universals According to Hockett	30
5.2	Universals According to Greenberg	31
5.3	Universals According to Chomsky	32
5.4	Other Definitions of Universals	33
5.5	Summary of the Definition of Universals in Language	37

The definition of “Universal” has been something that has occupied many thinkers throughout history and continues to do so today (Galluzzo & Loux, 2018; Sanford, 1980). Such considerations occur in philosophy, although they do not focus on language (Hierro Sánchez Pescador, 1989), but try to explain the general events of the world in the opposition between nominalism and universalism (Carmichael, 2010). In the case at hand, in this thesis, we work with Universals in language. For this reason, we do not go into detail on this type of philosophical proposal and refer the reader to the previous references since we must focus on Universals in language.

There is no unanimous definition of what is meant by Universals in Linguistics. Different authors have dealt with the concept, and different nuances can be appreciated in their respective definitions. This aspect was already highlighted decades ago: “In view of this development, it is surprising to find that little progress has been made in clarifying the central concept of language Universal” (Lieb, 1975:472). In general, we can group the definitions of Universals into three main groups, which refer to three entirely different definitions of linguistic Universals in the 1960s:

Hockett's, Greenberg's, and Chomsky's. Or, in other words, the features of human language as a communication system, the linguistic characteristics of certain independent language groups, and the properties of the generative grammar for a language, respectively.

5.1 Universals According to Hockett

“A Universal of language is a feature or property shared by all languages, or by all language. The assertion of a (putative) Universal language is a generalization about language”, according to Hockett (1963:1). First, we want to emphasize that the author understands that a rule is Universal if it is in 100% of languages. Secondly, Hockett draws an obvious link between a rule present in all languages and a generalization about language. We must understand this last expression as the intention to define language characteristics verified in the totality of the so-called languages.

Therefore, for Hockett (1963), formulating rules that manifest themselves in all languages allows for a definition of what language is. This means that we are giving the consideration of these Universals the power to help determine whether a new code is a language or not. We can point out that for Hockett, Universal is a synonym of his “design features of language”. This means that he is defining human language by detailing its various features. These are present in any natural language used by humans.

When we speak of the existence of different human languages in the world, we are assuming, *de facto*, two key aspects:

1. There are several common (Universal) features that allow us to determine that two different elements (or codes) are covered by the same concept (language).
2. We also assume the opposite: a series of specific features make each language different.

In other words, if we say that English is a language and Navajo is a language, they must share some (Universal) features of the category “language” that allow such a comparison. Nevertheless, there must also be differences that make us refer to two distinct languages and not just one. This vision is the one Pericliev (2012:212) shares also in his definition of Universal:

in saying that the languages of the world are diverse, we are already committed to the view that they possess properties that are common to all, and which allow us to call them ‘human languages’, that is, we are ultimately assuming the existence of Universals. On the other hand, the Universalist view, in claiming commonalities between the languages of the world, is committed to the view that these individual languages are distinct, that is, it ultimately assumes the existence of language diversity.

In short, Hockett is committed to the existence of characteristics shared by 100% of the languages that inherently underlie them. Therefore, he believes that one of the tasks of the linguist is to unravel all these latent Universals. He argues for considering that linguists should not “invent language Universals, but to discover them” (Hockett, 1963).

The implication of this determined commitment to Hockett’s Universals can already be seen in earlier reflections, such as that of Kemeny (1959), on treating exceptions. If we determine any rule as a proper and Universal characteristic of any human language (the ability to express tenses beyond the present, for example) and we discover a new possible language in the future that cannot comply with it, we should conclude that perhaps that discovered system is not a human language (Plank & Filimonova, 2000:100). This aspect, the naturalization of exceptions, is one of the most recognizable features of Joseph Greenberg’s proposal.

5.2 Universals According to Greenberg

For Greenberg, Osgood, and Jenkins (1963:255): “Language Universals are by their very nature summary statements about characteristics or tendencies shared by all human speakers”. This brief definition contains two key aspects that differentiate Greenberg’s proposal from the rest of his contemporaries:

1. He consciously speaks of “characteristics or tendencies”.
2. The shift of focus from language to speaker.

Regarding the first item, when defining something Universal, he adds the possibility that there is simply a majority behavior without being the only possible one. This aspect is the origin of the logical controversy behind using the label “Universal” for a crowded rule in 100% of the occasions.

We must point out that although the assumption of exceptions in Universals as natural is quite groundbreaking, it can be seen before Greenberg. Already in Bloomfield (1933:20) we find this fact described:

The only valuable generalizations about language are inductive generalizations. Features we think ought to be Universal may be absent from the very next language that becomes accessible. Some features, such as the distinction of verb-like and noun-like words as separate parts of speech, are common to many languages but lacking in others.

There is a commitment to the inductive method in Linguistics proposed by Bloomfield, which Greenberg will follow and discuss later. The main difference between the two proposals is that in Greenberg, the reference to the existence of exceptions appears more explicitly compared to Bloomfield.

Regarding the second item, we understand these words as a declaration of the intentions of the functionalism behind Greenberg's Universals. For Greenberg, the formulation of statements that are dominant in the various languages of the world indicates speakers' preference for a particular realization. These preferences, for Greenberg, will not be casual or accidental, in terms of Campbell, Albany, Bubenik, and Saxon (1988), and will have a basis in issues related to greater ease of processing, for example. In fact, according to Lieb (1975:472), the statements that Greenberg is trying to discover are "the major link between Linguistics and psychology and Linguistics and ethnology (or anthropology)".

5.3 Universals According to Chomsky

The third way of interpreting Universals is the one framed in the work of Chomsky (1965:28): "The study of Universals is the study of the properties of any generative grammar for a natural language. Particular assumptions about linguistic Universals may pertain to the syntactic, semantic, or phonological component, or interrelations among the three components".

This definition directly links the Universals with his proposal of generative grammar. This implies a direct link with the principles innately present in all humans because they are human beings. Therefore, any property present in this generative grammar is Universal for Chomsky. His formulation restricts it to only three components, which can be pretty restrictive. For this reason, some authors such as Cooreman and Goyvaerts (1980:621) point out the neglect of pragmatic Universals, for example.

For Hopper and Thompson (1993:136), "a Universal is a linguistic feature, usually phonological or grammatical, hypothesized as being shared by all human languages". Hopper's case serves us to see the tendency of linguists who understand Universals in a way close to Chomsky's to delimit the nature of linguistic features that can become Universals.

5.4 Other Definitions of Universals

Many authors have tried to offer their definitions of Universals over the years due to their disagreement with some of the aspects mentioned in the previous definitions. However, we must comment on two patterns of all the definitions found. On the one hand, we found none the scientific community has widely accepted as a solution to the definitional problems associated with the term Universal. On the other hand, any of the definitions found can be considered an evolution or update of any of the three visions presented above. In other words, no substantial changes can be observed to speak of a new way of understanding the phenomenon.

Comrie (1989:4), when proposing to define Universals understood as Chomsky, states: “Language Universals would be those innate linguistic principles which facilitate the child’s language-learning task”. In such a case, we see Universals as synonyms of principles (Epps, 2010:2).

The assimilation of the term Universal to principles can be seen again in later definitions by Chomsky himself, through publications by other researchers such as Cela-Conde and Marty (1997:27): “Within the framework of generative grammar, language Universals are taken to be properties of the initial state of the human faculty of language, not properties observed in all or most languages”.

In this case, “properties of the initial state of the human faculty of language” should be understood as a paraphrase of “principles”. The most striking fact about this definition, however, is the clarity and forcefulness with which it isolates and rejects a definition given by Chomsky himself of what a Universal is for a typologist.

We also find definitions in the opposite direction, i.e., from linguists who use Universals understood in Chomsky’s way as a counterpoint to indicate what is and what is not Universal, who understand Universals as cross-linguistically recurring patterns that emerge due to similarity in constraints and culture/history, but such patterns should be expected to be probabilistic tendencies, not the rigid properties of UG (Christiansen & Chater, 2008, 2009; Croft, 2009b; Evans & Levinson, 2009).

Other typologists such as Bickel (2010:77) propose more general definitions of Universals: “Language Universals are statements that are true of all languages”. The definition mixes the analysis of the world’s languages (typological methodology) with the restriction that the statement must apply to 100% of the cases analyzed. If this definition was strictly adopted, many of the claims

treated in the discipline would be left out, and the vast majority of implicational Universals would be discarded.

A different conception of typological Universals is proposed by Daniel (2010). His development posits the understanding of typological Universals as a co-variation of two parameters ranging from not being related at all to showing a strong tendency in their relation to full linkage. In other words, Daniel (2010) understands Universality as a continuum. However, the linguist merely points to this understanding of the phenomenon of the relationship of linguistic structures as a degree. There is no further development. This conception is interesting as a starting point for alternative classifications to Universals. It should be noted that Daniel's definition considers only the "co-variation of two parameters", i.e., implicational Universals of the type "if a language has X, then has Y". This view of Universality could also be extended to Universals where there is no relation between two parameters of the type "all languages have Y". The Universality value of this can also be conceived as a degree between the two extremes. The proposed conception of Universality as a degree is consistent with a proposal made more assertively in Typology in recent years concerning the degree of a given linguistic type that a language exhibits (Levshina et al., 2021).

Another of the definitions that we highlight positively in this type of Universals is the one proposed by Moravcsik (2010:1):

We will take a language Universal to be a grammatical characteristic that can be reasonably hypothesized to be present in all or most human languages. A Universal hypothesis is reasonable if it is based on a large, genetically, and areally balanced sample, if an independently motivated principle predicts it, or if both are the case. We will use the term 'Universal' both for the characteristic and the statement describing it.

Initially, it can be seen how it links the term Universal to any characteristic, regardless of its form. This contrasts with the precision (counterproductive, according to us) of the previous definition that spoke of the co-variation of two parameters. Another exciting element of the definition is that it reminds us that we always deal with a hypothesis. The characteristics labeled as Universal by Linguistics can hardly be contrasted in the totality of languages to which such a hypothesis is applied. Subsequently, the expression "all or most" is included, which makes it possible to capture both absolute Universals and implications with some exceptions. After the definition, Moravcsik offers us the requirements that the methodology used must have for the hypothesis to be reasonable, laconic, and complete. Finally, she resolves another of the reasonable doubts about which element is the Universal. It is pointed out that the designated characteristic and the statement are recognized as Universal. We share this view since we are dealing with a

metonymy that has caused us to call these statements as such.

This last part of Moravcsik’s definition is the one Matthews (2003:1783) uses exclusively to define what a Universal is: “a property that all languages have or a statement that holds for all languages”. Later, he adds that one can also find Universals that are not fulfilled in all languages but are not pure. Something similar to Crystal (2008)’s conception of Universals.

This same essence of Moravcsik appears in a more reduced but adequate definition in Pericliev (2012:212): Universals “are commonly conceived as propositions that are true for all, or almost all, languages (relative to a sample)”. As can be seen, the conception of the hypothesis is respected more indirectly by considering the formulations “propositions” that are also made “relative to a sample”. It is also important to highlight something present in most definitions that we believe is essential in the Greenbergian conception of Universals: “all, or almost all (languages)”. The most widespread and accepted is the understanding that a formulation tested in a representative sample of languages and with significant compliance (even if not total) is considered Universal. This definition will be maintained in general terms in subsequent publications of Pericliev (2018).

Christiansen, Collins, and Edelman (2009:4), focus, in the first place, on assuring that “at the present time, there is no need to justify the claim that language Universals exist”. Something in which Hornstein and Boeckx (2009:79) also agree: “What the Universals are is open to debate, even though their existence is, in our opinion, not rationally contestable”. This is because, regardless of one’s conception of what a Universal is, all approaches have demonstrated satisfactory results. Christiansen et al. (2009:4) do not make invisible that “opinions differ sometimes strongly over the exact nature of language Universals, their origin, and how best to study them”. Such a publication is an example of a common framework for studies of the Universals of the three conceptions presented here.

In this sense, we find other definitions that try to reconcile and bring together the different conceptions of Universality that have been shown. Siemund (2011:1) indicates that:

Linguistic Universals ideally are generalizations that capture properties of language or languages that are essential to and stable across all possible languages and language types. In principle, they even extend to languages that are not attested, and hence also apply to languages that are extinct or have not yet come into existence. Linguistic Universals also define the scope within which languages can vary and make predictions concerning the co-occurrence of structural properties.

The presence of “ideally” can be misleading. That is to say, it is synonymous with what Universals should aspire to but is not a rather problematic fact. This is to include reference to language in general as well as to languages to capture those different definitions discussed above.

We also note another duality in the contrast between “essential” and “stable”. We can understand that “essential” is something that must be in any language because it is a language. Stability is attributed to the different linguistic types, with a reference closer to Greenberg’s perspective. Finally, something little seen in the previous definitions is shown: the power and extension of Universals.

Another example of Greenberg and Chomsky’s unification of Universals can be found in Garrette (2015:6). The researcher states that:

the concept of Universal grammar, in the strictest sense, refers to a theory promoted by Chomsky(1965) which argues that all natural languages have a common underlying structure, and that the human brain is hard-wired with a small set of grammar rules that give us the ability to learn language by mapping what we hear onto that structure. For our purposes, we use a weaker definition of Universal, in which we recognize that, cross-lingually, languages tend to share certain broad grammatical properties.

We wanted to highlight this definition of Universal because of his conception that there is a strong definition (oriented to Chomsky) and a weaker definition (oriented to Greenberg). For Garrette, having no exceptions is simply a stricter requirement, while accepting exceptions is a relaxation of this requirement. However, both approaches to the concept of Universal could be equally strong or weak. The presence of exceptions is only due to the characteristics of the rules analyzed, which differ significantly from one methodology to another (typological Universals are usually more specific, which facilitates the presence of exceptions).

There are also differences in the characteristics of the analyzed rules between the Universals understood by Chomsky and by Hockett, for example. This is something that is perfectly explained by Wacewicz and Żywicznyński (2015:37):

These shortcomings, as viewed from the perspective of language evolution, hinge upon two (closely related, but distinct) major issues: firstly, Hockett’s classificatory system focuses on the means, especially the physical properties of the medium of transmission, at the expense of content, and secondly, it focuses on the code itself, rather than the cognitive abilities of its users that make language use possible in the first place

However, such a difference does not usually influence the presence of exceptions to which the formulated rule applies. In both cases, we usually find rules present in all languages. An example of a review to provide exceptions to the work of Hockett (1963) can be seen in Blasi, Wichmann, Hammarström, Stadler, and Christiansen (2016).

The best explanation, in our view, of both the origins and the consequences of the different views of the Universal concept is by Croft (2003:4-5):

The generative approach represents a reaction against behavioristic psychology, while the typological approach represents a reaction against anthropological relativism. [...] The behaviorist view of language, in particular language learning, is anti-Universalist in that it posits no innate, Universal internal mental abilities or schemas. In the behaviorist view, linguistic competence is acquired through learning of stimulus-response patterns. In contrast, the generative approach posits the existence of innate internal linguistic abilities and constraints that play a major role in the acquisition of language. [...] The anthropological relativist view of language is that languages of the world can vary arbitrarily: ‘languages could differ from each other without limit and in unpredictable ways’. [...] Greenberg and others discovered that a more systematic sampling of a substantial number of languages reveals not only the range of variation but constraints on that variation. Those constraints demonstrate that languages do not vary infinitely, and the constraints represent linguistic Universals

In this quote, we highlight, above all, a very significant element not previously discussed. The Chomskyan view of Universals focuses on finding essential constraints in language acquisition (and therefore present in any language). The Greenbergian view of Universals pursues the systematization of constraints in variation.

5.5 Summary of the Definition of Universals in Language

To sum up, we have seen multiple definitions for the same label. We believe that a good way to summarize graphically the different ideas that linguists have in mind when they think of Universals is to rescue the comparison established by Dryer (1998:2), in the figure 5.1:

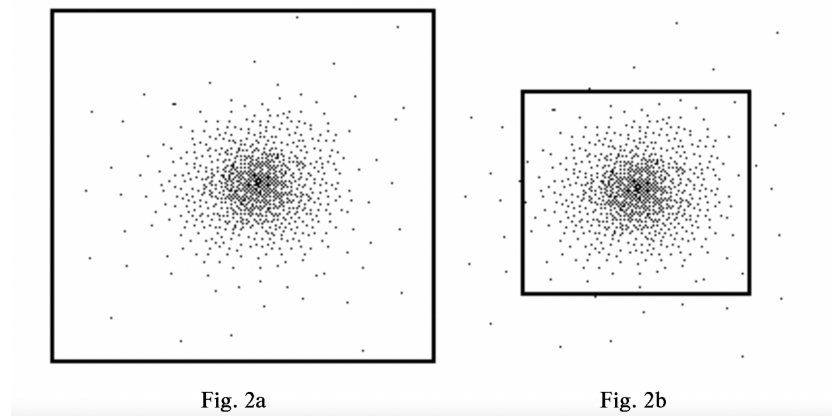


Figure 5.1: Comparison between a Conception of Universal as the Possible and Universal as the Probable in Dryer (1998).

In other words, two possibilities are represented:

- In the square on the left, we can see the interest in delimiting the possibilities of human languages.

- In the square on the right, we can see the interest in delimiting the tendencies of human languages.

While in the first case, it is a question of formulating limits that allow them to gather within them all the languages of the world, in the second case, it is a question of formulating limits that allow them to gather within them the clear behavior of the languages of the world. The interest of this square on the right is that it is much more precise despite leaving some points outside the square. In Linguistics, we can translate this into more meaningful or detailed claims. However, we must emphasize that Linguistics requires the knowledge offered in both squares. Having the different information contained in them is vital, and we do not believe them to be mutually exclusive (Smith & Tsimpli, 1995). As can be seen, the only problem is that they share the same label, and this needs to be clarified.

The problem we find with the definitions of Universal lies in how Universality is understood and what one wants to achieve. One of the most widespread solutions to the problem is the use of different labels, specializing the concept of Universal for one of them and creating new expressions to capture such differentiation, as reflected in its definition in Grandi (2020:75):

i primi (Universals) indicano quelle proprietà, correlazioni o strutture linguistiche che, senza alcuna eccezione ricorrono in ogni lingua storico-naturale. Le seconde (tendencies) designano le proprietà, le correlazioni o le strutture linguistiche che sono attestate in una porzione statisticamente rilevante delle lingue storico-naturali.

Once we reviewed the different proposals for defining Universality, we believe it is useful to highlight the main points of the main visions in table 5.1:

Characteristic	Hockett	Greenberg	Chomsky
Linguistic explanation of human nature	+	+	+
Rules of human communication system	+	-	-
Tendencies of some human languages	-	+	-
Grammar part shared in all human languages	-	-	+
Allows exceptions	-	+	-
Allows clustering	-	+	-
Mainly in word order	-	+	-
Mainly in conditional form	-	+	-
Possible in syntax, semantics, and phonology	-	-	+
Based on a relatively big sample	-	+	-

Table 5.1: Differences in the Different Views of Universals in Language

6

Terminology Used in Universals of Language

Contents		
6.1	Labels to Distinguish the Conception of Universal	40
6.2	Labels to Distinguish the Structure of the Universal	47
6.3	Labels to Distinguish the Presence of Exceptions	51
6.4	Labels that Are Used for Different Variables	61

In the literature dealing with Universals, one can observe many labels accompanying *Universal*. Therefore, we have decided to analyze, sort, and quantify this. As a result, we have collected about 60 different expressions related to *Universal*.

This large number of different elements that we have found can be gathered into three main groups about the key aspects that they are intended to underline:

1. Differences in the conception of Universal and its research methodology.

This point involves the different conceptions of what a Universal is in Linguistics.

2. Differences in the structure of the Universal.

The differences in the structure respond to a formulation of the Universal restricted by a precondition or without any condition.

3. Differences in the presence of exceptions.

We can find Universals with or without proven counterexamples.

4. Labels that are used for different variables.

[illegible]

6.1 Labels to Distinguish the Conception of Universal

- Aristotelian explanations would resemble the work developed by Greenberg: formulations based on observable data through statistics, with the acceptance of exceptions.
- Galilean explanations would resemble the work developed by Chomsky: formal and abstract formulations, without the presence of any exceptions (as it happens with Hockett, too).

The use of these two labels, with a philosophical basis, is the embodiment of the two general currents known as empiricism (comparable to Aristotelian explanations) and rationalism (comparable to Galilean explanations), as discussed by Bach and Chao (2009); Cooreman and Goyvaerts (1980). Boeckx, Fodor, and Glertman (2009) use the distinction between Aristotelian and Galilean Universals, too.

Other authors prefer to establish the distinction between e-Universals and i-Universals (Hornstein & Boeckx, 2009). In this case, the “e” refers to “externalist” and the “i” to “internalist”. That is:

- I-Universals correspond to formulations with an accent on innatism and internal rules shared by any human.
- E-Universals correspond to formulations emphasizing the characteristics observable in the different languages of the world and their invariances.

In fact, so strong is the weight of innatism on Universals understood as Chomsky that Bybee (2009:22) proposes the label “innate Universals” to distinguish them from Universals understood by Greenberg. Specifically, she points out that:

As mentioned, in Chomskian theory, Universals emerge in child language acquisition. In two senses, this approach has yet to be fully articulated. First, despite some similarities between preferred patterns in the languages of the world and early child language, strict correspondences, whether phonological, morphological, or syntactic, have been hard to locate. Indeed, trying to locate language Universals empirically is a minor part of the generative research agenda. Instead, innate Universals are taken as given, so any feature demonstrated to be present in a few languages can be considered Universal.

We consider especially relevant the last sentence of the quote we have rescued, where it explains how a demonstration of a characteristic present in “few languages” can already be considered Universal in such an approach, thanks to the relation that these characteristics have with innatism (Mendívil-Giró, 2018).

Cristofaro (2010:2) criticizes the lack of evidence, something that, according to her, does not exist when analyzing concrete realizations of different languages:

there is no evidence that a speaker’s linguistic knowledge consists of an inventory of Universal grammatical categories and relations that can be defined formally. Contrary to what is assumed by generative linguists and many typologically orientated linguists, we only have distributional evidence for language-specific and construction-specific grammatical categories and relations

Bickel (2017b); Ibbotson and Tomasello (2016); Lin (2017); Siemund (2011) agree with the above quote, emphasizing, even more, the functional motivations (processing, time or memory constraints, etc.) behind Greenberg’s conception of Universals compared to Chomsky’s research method.

As can be seen, the distinctions and the various labels always revolve around the Greenberg-Chomsky opposition. No characterizations of Universals understood as Hockett’s are offered. This is because such formulations, soon after being published, ceased to be recognized under the Universal label and became known as “design features of human language”. This change (and establishment of the term) means there is no need for any nuance regarding Universals. However, in the bibliography, references to the Universal label can still be found in Hockett’s work, a fact that requires some extra characterization, such as: “definitional Universals” (Plank & Filimonova, 2000). In the case of Newmeyer (2007:16), he uses the label “features of language” for definitional Universals.

However, the labels that we observe that enjoy the most success (only because they are seen more frequently) are:

- Deep Universal
- Surface Universal

The main difference in order to justify the use of these terms lies in the type of properties being analyzed (Newmeyer, 2008). Surface Universals (Greenberg ones), according to Newmeyer (2008:35), are based on an:

examination of large samples of languages, extracts inductive generalizations by realizing easily identifiable grammatical properties of the languages in the sample, and further generalizes the results to human language as a whole. For the most part, participants in this tradition seek explanations outside of grammatical theory per se for the generalizations they discover. In particular, most adopters of the surface Universals approach hold to some version of functional or cognitive Linguistics. However, many publications in this approach are content to leave the Universals they propose unexplained, or at least fail to see the task of description and explanation that there is the sense that providing the former without the latter results in an ‘incomplete’ presentation.

In contrast, for Newmeyer (2008:36), deep Universals (Chomsky ones):

are practiced almost exclusively by proponents of one or another variety of generative grammar. On this view, Universals are properties of Universal Grammar (UG) and, thus, a reflection of our innate biological endowment for language. Universals in this sense (hereafter “deep Universals”) are thus typically highly abstract entities that cannot be deduced inductively from examining the surface properties of the languages in a sample, whatever the size of

the sample. Generativists do not reject the existence of easily identifiable cross-linguistic tendencies, of course, but they regard the task of discovering such ‘surface Universals’ as secondary, since the task does not necessarily facilitate the more primary one of elucidating UG properties.

Comparing the above two quotes, one can see multiple details that differ between deep and surface Universals. They can all be summarized in that the former are “theory-driven Universals” and the latter is “data-driven Universals” (Comrie, 2006). This difference in the data on which the claims are based is critical to understanding the elements under review and the methodology followed. The surface Universals approach can be compared to the natural science approach (as opposed to the deep or theoretical approach) in the terms presented in Itkonen (1980a). The data that are usually treated in deep Universals usually correspond to what Corbett (2009) encompasses in “feature” (gender, number,...) or even in more abstract entities. The data that are usually treated in surface Universals usually correspond to what Corbett (2009:130) encompasses in “value” (feminine, neuter, dual, plural,...).

As can be seen, the properties analyzed are superficial, derived inductively in a sample of languages in the case of surface Universals. In contrast, deep Universals are abstract entities related to the properties of Universal Grammar. The sample must be large and significant in the case of surface Universals, while Coopmans (1983:455-458) shows how a single language is enough for deep Universals claims. Alternatively, in other words, as pointed out by Polinsky (2010:8): “one orientation refrains from pushing for deeper generalizations, while the other desperately tries to derive every observable phenomenon from structural principles”.

Another possibility in distinguishing deep and surface Universals that is relatively widely observed in the literature is the distinction between:

- Universals of Language
- Linguistic Universals

Although not all authors make firm use of such a distinction, we see how even Chomsky (1965) himself constantly references “linguistic Universals”, the statements he formulated. Thus, the linguists closest to Greenberg usually speak of “Universals of language”, which is, in fact, the name he uses for his publication (Greenberg, 1963).

In our opinion, the denomination used by Abraham (1989) is a more transparent way of doing this. In this case, the surnames of the founders of the approach are used to show the difference, giving rise to:

- Greenbergian Universals
- Chomskian Universals

This same author even uses the expression linear Universals at some point as a synonym for Greenbergian Universals. Linear should be understood as a synonym of “word order”. However, we must point out that the correlation is not exact. Not all typological Universals are based on word order, and not even all of Greenberg’s Universals refer to such a word order. However, most of them indeed are. Therefore, this would be a generalization.

Another proposal is made by Odden (2003). First, it contrasts definitional Universals with observational Universals, a label also appearing in Pericliev (2012). The former would refer to Hockett’s proposals, and the latter would include the proposals of Chomsky and Greenberg. Subsequently, he makes a distinction between two types of definitional Universals, which he calls “true Universals” (Chomsky’s) and “Universal tendency” (Greenberg’s), a distinction also done by Dabrowska (2015). We believe that such an influential label as the adjective “true” undermines the work of Linguistic Typology and can also be misleading. Another contrast we find a few pages later in the same reference is in the labels “architectural Universals” and “observational Universals”. The former would correspond to the claims proposed by Chomsky and the latter to claims proposed by Greenberg.

In the same 2003, we find two publications that use different labels to refer to surface Universals:

- Cross-linguistic Universals (Croft, 2003:87).
- Typological Generalisation (Cysouw, 2003:89).

However, Levshina (2022:13) prefers to merge them in “cross-linguistic generalization”. We find this type of nomenclature more accurate and transparent. In the first case, reference is made to the key and distinctive aspect of the methodology of this type of approach, based on the cross-referencing of data from different languages and their comparison. In the second case, reference is made to the sub-discipline that carries out this type of Universals and uses the term “generalization”, which we see as a good way of capturing many of the data seen above in the definitions of Universal. It especially emphasizes the character of hypothetical rule that wants to capture some invariances in most languages. The term cross-linguistic contrasts nicely with Croft (2009a:145)’s definition of the deep Universals methodology, which he characterizes as “one

language at a time approach”. Whaley (2010) uses the term “cross-linguistic patterns”, similar to the previous ones.

Moreover, we have to mention that, in the case of deep Universals, Chomsky (1965:28) goes further and presents a classification of these through two other labels: “A theory of substantive Universals claims that items of a particular kind in any language must be drawn from a fixed class of items. [...] The property of having a grammar meeting a certain abstract condition might be called a formal linguistic Universal”. Further development of these two elements beyond their definition is not necessary because they are not the object of interest of this work. Nevertheless, it can also be extended in Cooreman and Goyvaerts (1980); Dabrowska (2015).

Some authors, such as Guevara and Scalise (2009), believe that the debate between Chomskian and Greenbergian approaches remains unresolved and that this distinction will be necessary for a long time. They believe there is no room for a common field. Such a position contrasts with other contemporaries such as Newmeyer (2008), Croft (2009a), or Dryer (2019), who mention the mutual benefits of both approaches and how they are complementary. For example, Croft (2009a:148) perfectly captures the commonalities in both approaches:

This conclusion leaves us with two of the views outlined above. Formalists and functionalists are examining the same phenomena. Both are positing grammatical categories and grammatical distinctions supported by linguistic behavior. Both use a range of distributional facts to support their analyses of grammar. Nevertheless, the result, the language Universals proposed, is very different. The generative, and more broadly formalist, Universals of syntax are Universals of structural properties of linguistic form that constrain possible formal grammars (the “principles” of more recent generative grammar; other formal theories such as Head-driven Phrase Structure Grammar and Lexical Functional Grammar have similar principles). The typological Universals of syntax are Universals that constrain how functions are expressed in linguistic form. Why is this so? In the next section, I argue that it is because the methods of syntactic analysis are fundamentally different in the two approaches.

Besides the term “typological Universal” in Croft (2009a), we have also widely observed it, especially in recent years. Researchers such as Cristofaro (2019) or Guzmán Naranjo and Becker (2022) use it.

Christiansen et al. (2009:4) also agree in viewing Greenberg’s approach and Chomsky’s as interrelated. They state that:

Although the origins of these two streams of thought are very different, the dividing line is becoming less clear. In an attempt to delineate possible parametric variation (and hence to isolate properties of Universal Grammar), generative work has increasingly focused on a wide variety of typologically very different languages. The interpenetration and dialogue between these two streams of thought is one of the defining characteristics of the Linguistics of the second half of the twentieth century and the beginning of the twenty-first century.

This statement exemplifies that differences are increasingly blurred, as evidenced by the type of languages chosen for generativist analyses. In other words, English is (relatively) less and less used, although we prefer to develop this issue further when we deal with Anglocentrism in Linguistics, a fundamental fact for studies on Universals.

In other words, we believe that a good summary of what is involved in the creation and use of the labels “surface Universal” and “deep Universal” to reflect two distinct realities is nicely summed up in the words of Comrie (1989:1-2):

The two approaches can be contrasted on a number of parameters, the most important of these being the following: the data base for research on language Universals (a wide range of languages, or a highly restricted set of languages); the degree of abstractness of analysis that is required in order to state language Universals (for instance, in terms of more concrete or more abstract syntactic analysis); and the kinds of explanations advanced for the existence of language Universals. [...] On the one hand, some linguists have argued that in order to carry out research on language Universals, it is necessary to have data from a wide range of languages; linguists advocating this approach have tended to concentrate on Universals stable in terms of relatively concrete rather than very abstract analyses, and have tended to be open, or at least eclectic, in the kinds of explanations that may be advanced for the existence of language Universals. On the other hand, some linguists have argued that the best way to learn about language Universals is by the detailed study of a small number of languages; such linguists have also advocated stating language Universals in terms of abstract structures and have tended to favor innateness as the explanation for language Universals.

We want to highlight the connection behind the labels we have shown. Although in the beginning, as we have seen, there were evident boundaries between surface Universal and deep Universal, a few years later, a rapprochement began to emerge (Cooreman & Goyvaerts, 1980:616). This view is also shared by Hawkins (1980) in the same year. We understand that approaches to the same phenomena with different methodologies can be exciting, as they allow us to confirm or qualify the different hypotheses.

An interesting proposal is the one used in a paper by Masini and Mattioli (2019), which is called the “multi-level approach”. This consists of studying a linguistic phenomenon through different methodologies and data sets. A more in-depth study of one or two languages is not incompatible with a more superficial study of many languages. The data can help to gain a multipolar understanding of the phenomenon.

There are many labels to distinguish the typological Universals (the ones we are working on in this thesis) from the rest. Many labels used for the same thing make the work and understanding complex, especially for those less familiar with them. The terminological variety observed can sometimes even lead to the inappropriate use of the different labels.

6.2 Labels to Distinguish the Structure of the Universal

In Greenberg (1963:x)’s original proposal, there is already a particular interest in the formal structure of the Universal that can be formulated. Even in the preamble of his foundational work, there is a distinction between:

- Simple Universals
- Implicational Universals

The former refers to rules that apply to all the languages of the world (in all languages Y). The latter refers to rules that apply to languages with a particular condition (if a language has X, it has Y).

Actually, Greenberg (1963:58) goes much further as to assert that “what seem to be non-implicational Universals about language are tacitly implicational since they are implied by the definitional characteristics of language”. In such a case, Greenberg has already changed the label “simple Universals” to “non-implicational Universals”. The exciting value of an implicational Universal is the reading it makes in which a given feature seems to be the direct trigger of another feature. In the case of the “definitional characteristics of language”, they are not any direct trigger of other characteristics but are essential elements and are, therefore, not fundamental in assessing the behavior of other elements in a concrete way. In other words, we believe that the mere fact of posing an implication of the type “if a language puts the nominal object after the verb as the dominant order, then a language has displacement (allow to talk about things not physically present or that do not exist)” is useless (Greenberg, 1963:58). We will not find cases contradicting this implication (languages that put the nominal object postverbal and have no displacement). However, such an implication is useless because one variable is not related to the other. This proves that languages that preverbally place the nominal direct object exhibit the same attitude.

Our view is the same as Pericliev (2012:213). This researcher, speaking of other cases of misformulation of Universals, comments that:

Also, one can truly assert that if one eats some food (P) he will die at some future time (Q), but this true implication $P \Rightarrow Q$ will be misleading in suggesting that the food is the actual cause of the death, whereas the empirical truth of ‘everyone will die or humans are mortal’ is the real reason why the whole statement ‘if one eats some food (P) he will die at some future time (Q)’ will necessarily be true (logically, Q implies $P \Rightarrow Q$).

Another example in this sense is given by Hawkins (1983:85) with the example: “*Postp* $\Rightarrow ((AN \vee NA) \wedge (RelN \vee NRel))$ ”. In such a formalization, it is said that languages that tend to have postpositions exhibit adjectives before the noun or after it and relatives before the noun or after it. Such an implication, which encompasses one type and its opposite, does not provide relevant information for characterizing or grouping languages. In any case, it is necessary to go beyond implications that are structurally correct in logic but useless in Typology.

Greenberg et al. (1963:258) later refer to a label in which the head and dependent of the expression are changed: “Universal implications”. In any case, this refers to the same thing. Four logical possibilities are being mentioned, and one is said not to be possible in the languages (or it is residual). Hawkins (1980:200) explains this by the following basic formalism:

“If P then Q $>$ (* P & -Q) but yes (P & Q) (-P & Q) (-P & -Q)”.

We must understand it as if we are pointing out that if P exists, then Q exists. This implies that we have four logical possibilities, three of them valid. It is possible that:

- P and Q exist.
- P does not exist, but Q does.
- P and Q do not exist.

It is not possible that we find the situation in which:

- P exists, but Q does not exist.

In natural language and using a concrete example, we can say that the Universal implication “if a language has the dual number, it has the plural number” encompasses the following possibilities:

- A language has dual number and plural number.
- A language has neither dual number nor plural number.
- A language has no dual number, but plural number.
- A language has dual number, but no plural number.

Of the four possibilities shown above, only the last one is not to be found in the world’s languages, according to Greenberg. Alternatively, in the words of Croft (2003:53): “Implicational

Universals differ from unrestricted Universals in that they do not assert that all languages belong to one type. Instead, they describe a restriction on logically possible language types that limits linguistic variation but does not eliminate it”.

Such a structure for formulating Universals, although it may seem more limited (as it applies to fewer languages), is much more interesting for typologists, as Hawkins (1979:618) demonstrates already in the title of his paper: “Implicational Universals as predictors of word order change”. Languages that are in constant flux may eventually acquire a different linguistic type. If that were to happen, the Universal would allow us to know the next movement of this (the other part of the implicational Universal).

This interest can be exemplified in the quote from Croft (2003:54): “What makes implicational Universals more interesting than unrestricted Universals above all, is that they state a dependency between two logically independent parameters”. Comrie (1989) also uses the opposition between the labels “implicational” and “non-implicational”. In the case of Plank and Filimonova (2000), they prefer to use the terminology of “unconditional” and “conditional”, although the conception is the same. This is consistent with Greenberg et al. (1963:258)’s original proposal, although in that case, the word “unrestricted” was used as the equivalent of “unconditional”.

Levshina (2022:13) chooses to use the term “typological implication”. This kind of transparent nomenclature has both positive and negative points. It is interesting to show clearly the relation between typological Universals and implications. On the other hand, we believe it can present problems when discussing cases contrary to the “typological implication”, which Levshina (2022) does not mention. Moreover, linking implications directly to typological Universals makes it more challenging to fit less canonical or less frequent, but technically possible, structures such as those without a conditional structure.

We should also mention another type of implications that are much more residual and specific. That is terms such as “restricted equivalence” (Greenberg et al., 1963). Here, we refer to a mutual implication of two unusual language features. For example, if a language has a dental click, it has a lateral click, and vice versa. In other words, it is possible only to find two manifestations in languages:

- A language has dental clicks and lateral clicks.
- A language has neither dental clicks nor lateral clicks.

That is, the following two logical possibilities cannot be attested, according to Greenberg:

- A language has dental clicks, but no lateral clicks.
- A language has no dental clicks, but lateral clicks.

Another rather specific but critical label in the Universals presented by Greenberg is what Croft (2003:58) names “biconditional Universal”. In this case, it is a double implicature involving two antagonistic types with antagonistic results. Levshina (2022:12) prefers to call them “typological correlation”. The only difference in the latter case is that it is more restrictive since it assumes that a double implicational Universal such as the one mentioned above can only be given in terms of Linguistic Typology and not in Universals of another type. A double implicational would be a structure found in surface Universals.

An example of this would be when Greenberg (1963) formulates that: “In languages with prepositions, the genitive always follow the governing noun, while in languages with postpositions it almost always precedes”. As can be seen, one direction is assigned to the genitive for a particular order of the adposition. In contrast, the opposite direction of the genitive is assigned to the opposite order of the adposition. This double implication of contrary types shows even more clearly that if it is a valid Universal, the element selected as the trigger of another direction (in this case), although logically independent, has an evident influence.

The implicational hierarchies are the last terms that qualify or complete the implicational Universals. Both Fox (1987) and Luo (1994) define it as an “implicational scale”. This definition may be laconic. Therefore, an implicational hierarchy is a sum of implicational Universals whose order cannot be altered. The appearance of an element in a language implies the appearance of all the elements that appear previously on the scale without being able to omit any of them. For example, as in the implicational hierarchy of Berlin and Kay (1969) in figure 6.2.

$$\begin{bmatrix} \text{white} \\ \text{black} \end{bmatrix} < \text{red} < \begin{bmatrix} \text{green} \\ \text{yellow} \end{bmatrix} < \text{blue} < \text{brown} < \begin{bmatrix} \text{purple} \\ \text{pink} \\ \text{orange} \\ \text{gray} \end{bmatrix}$$

Figure 6.2: Implicational Hierarchy from Berlin & Kay (1969)

That is if a language has a specific label to refer to what is understood in English as “blue”, then it will also have an equivalent for “green”, “yellow”, “red”, “white”, and “black”. Furthermore, if the language includes any more labels in the future, it will be “brown” according to the implicational hierarchy, while if it loses any, it will be “blue”. To expand on this concept, with other illustrative

examples, we recommend reading Siewierska (1988) and Allan (1987).

In this case, we observe less disparity of labels to refer to the structure of Universals. Combining “implicational Universals” and “unrestricted Universals” labels is usually the most accepted and straightforward. It is also necessary to point out the precise relationship between implicational Universals, typological Universals, and unrestricted Universals with the other type, as we will develop further below. However, it is also necessary to clarify that such a statement can only be generalized since, technically, there are other types of combinations between Universal modality and structure (although they are less frequent). For this reason, it is necessary to develop the two structures of formulating Universals, whatever their type, in an aseptic and independent way.

6.3 Labels to Distinguish the Presence of Exceptions

The last parameter that generates many labels is the presence or absence of exceptions in the formulated linguistic hypothesis. Common sense invites us to think it is a contradiction to speak of a feature as Universal but not in all languages.

The first defenses of Universals with or without exceptions and their nomenclature are already found in Jakobson (1963:212), who differentiates between:

- Exceptionless Universals
- Statistical Uniformities

The linguist speaks of a strictly probabilistic level and comments that the significance of this type of statement does not change: “Statistical uniformities with a probability slightly less than one are no less significant than uniformities with a probability of one”. Moreover, in this author, we also see the appearance of the term “near-Universal” as a synonym for “statistical uniformity”, something we also see in Greenberg et al. (1963). There is also the possibility of distinguishing between “near-Universal” and “statistical uniformity”. In this case, when we use the label “near-Universal”, we would highlight fewer exceptions than when we speak of “statistical uniformity”.

The mention of probabilities is something that is also observed in the descriptions of the Universals of Greenberg et al. (1963:259): “For any language, a certain characteristic has a greater probability than some other (frequently its negative)”. The American linguist further exemplifies this with nasal consonants: “The probability of a language having at least one nasal consonant

is greater (in this instance far greater) than that it will lack nasal consonants” Greenberg et al. (1963:259).

Since Universals are formulated in natural language, expressions that capture the different degrees of probability expressed in the different Universals formulated are needed. The expressions that Greenberg et al. (1963) add to Universals for this purpose are famous: “with overwhelmingly greater than chance frequency”, “almost always”, or “always”. All these expressions present in many Greenberg Universals can be called “Universal quantifiers” according to Plank and Filimonova (2000). The problem with the Universal quantifiers used in Greenberg is that there is no standardization of the values contained in such expressions, nor is their use justified to the detriment of the other.

Osgood (1963) calls them “non-Universal”. This strong label, however, does not correspond to a rejection of this type of Universals, quite the contrary, since he qualifies them as “the most interesting to a science of language behavior” (Osgood, 1963:239). These various defenses of Universals that do not apply to 100% of the languages analyzed are based on the idea that “previous endeavors to define the interlingual invariants in absolute metrical terms could only fail” (Jakobson, 1963:209), something that is also supported by Cooreman and Goyvaerts (1980) when they talk about the interest in a “near-Universal” being just as significant as that of an “absolutely Universal”. That is, when analyzing the invariance of the realizations of the world’s different languages if we set the criterion of acceptance of a claim to 100%, we will have practically no results. We must remember that this is because, in this type of test, we do not usually check common characteristics in the languages of the world but rather regularities in the variation of the world’s languages.

Therefore, the treatment of exceptions is subject to the conception of Universals. Their naturalization will be extensively discussed in typological-functional proposals. A summary of this differentiation is made by Levinson (2011:653):

With the rise of GENERATIVE GRAMMAR, and the insistence on Universals of grammar (see Universal GRAMMAR), anthropologists began to emphasize the diversity of language use, implicitly accepting the underlying uniformity of grammar (Hymes 1982). But with the growth of linguistic Typology and the empirical search for language Universals, it has become increasingly clear that real Universals -in the straight-forward sense, properties that all languages have- are vanishingly rare (at least beyond the basic organizational principles outlined by Hockett 1960, and some of the architectural properties sketched by Jackendoff 2002). Instead, linguistic typologists have found that empirical generalizations are nearly always of the kind “Across all languages, if a language has property X, then it probably also has property Y”. Meanwhile, generative grammarians have hoped to account for the diversity in terms of a limited set of variants (see PRINCIPLES AND PARAMETERS THEORY), but such variants are not manifested in grammars in any straight-forward way.

[Emphasis on capital letters present in the original quotation.]

Greenberg (1978b:43), a decade later, shows disagreement with the terminology previously used by him and his colleagues: “It might be advisable to abandon the current term ‘Universal’ as being misleading and substitute for it some such expression as cross-linguistic generalization”. The reasons for this are related to what he considers, then, a lack of correspondence between the original meaning of the term “Universal” and the type of Universals he formulates (implications that try to represent absences of occurrences of specific combinations of orders). However, the term Universal will already be very well established and naturalized, which will cause such a reflection to be little considered. Such an expression will be taken up many years later by Miestamo, Bakker, and Arppe (2016:235) to speak of “cross-linguistic tendencies”. These two expressions are equivalent since generalization and tendency capture the essence of the majority adoption of the world’s languages very well.

The functionalism that prevails in Linguistic Typology studies explains the emergence of Universals and the emergence of exceptions within these Universals. We find fascinating the explanation offered by Campbell et al. (1988:221) when they state that: “Universals hold best when internal factors (of production and perception) are at work, but that some Universals can actually have exceptions when external, socio-cultural factors are at play”. In this case, the behavior of languages is seen as a drive of factors favoring uniqueness or disparity between languages. When the weight of factors (extralinguistic according to the previous quote) favoring disparity is very low compared to the strength of factors favoring uniformity, or there are no such factors, such a Universal will not have exceptions. This is usually known as “interference” (Muysken, 2000; Thomason, 2001). That is, according to Koptjevskaja Tamm (2010:2), an

import or transfer of linguistic features from one language to another, with various possible modifications of the imported feature during this process. In borrowing situations, speakers maintain their native language but incorporate features of another language into it.

Thus, for Campbell et al. (1988:222), “in such cases, a proposed Universal does not cease to be a Universal simply because it may have a very few externally-determined exceptions, but rather, it becomes a statistical Universal”. From these words, we understand that the power of the cause that generates the statistical Universal is the same as that of an absolute Universal. The only difference is that statistical Universals have, in addition, an external cause in the opposite direction that can influence some of the languages of the analyzed world. Consequently, Campbell et al. (1988:222) conclude that

The important questions all remain the same. We still want to know why (for what perceptual, physical, psychological, biological, or pragmatic reasons) the trait occurs in all or nearly all languages, and why only one or a few languages depart from the otherwise Universal norm.

Perhaps the degree to which a Universal can have exceptions (i.e., be statistical rather than absolute) is correlated to the strength of its underlying explanations.

Furthermore, it should be noted that Campbell et al. (1988) opt for the use of the expression that has subsequently been more widely used and assumed in Typology studies: “statistical Universal” (Christiansen et al., 2009). This refers to the statistically significant value of the linguistic hypothesis proposed, even though its value is not equivalent to the maximum value.

In fact, Campbell et al. (1988:222) insist again in their paper that: “We are interested in general principles of language, in how the pieces of languages interrelate, how they function, and how they may change. Some of these principles have been considered to be absolute Universals, others near- (or statistical) Universals”. Thus, they again equate absolute Universals with statistical Universals, as both manifest different tendencies exhibited by the world’s languages. Like Merlo (2016:19), who also recalls that “the most frequent word orders have lowest cost of derivation or lowest deviance from some general principle of grammatical organisation”.

We have encountered an author who makes a different use of the label “statistical Universal”: Tanaka-Ishii (2021). In this book, this label is applied to the analyzed Universals. However, these Universals are statistics from processing large amounts of data embodied in mathematical functions. In short, it is the way he uses to name aspects such as Zipf’s Law (Zipf, 1935). Greenberg et al. (1963:260) call them “Universal frequency correlations”. The book of Tanaka-Ishii (2021), moreover, denounces the problem still existing at the date of its publication of a failure to determine the languages that may not adhere to the Universal posited to further determine that property as Universal.

Comrie (1989) offers one more comparison example to understand why the presence of exceptions in Linguistic Typology, related to a concrete and banal aspect of the English language, is naturalized. He comments that in descriptive Linguistics, there are laws, such as the formation of the past tense in English. A general rule allows us to apply it to new lexical items to which it has never been applied before, and we would do so satisfactorily (adding -ed to the verb root). However, we know of cases with exceptions to such a rule. This does not invalidate the rule of past tense formation in English.

Comrie (1989) opts for the label “nearly exceptionless Universals”. In this case, it is too long and complex to be used. Moreover, in the case of both Comrie (1989) and Campbell et al. (1988), who defend the exact value of Universals with the distinction between with or without exceptions, we believe that they would be more consistent if they did not propose any terminological distinction.

That is, returning to the example of the English past tense formation rule, it would be the same as if it had a different nomenclature (such as “imperfect rule” or “statistical rule”) to distinguish it from other grammatical rules of the English language that do not exhibit any exceptions (which will be very few, by the way).

Dryer (1989:283) prefers to use the term “significant statistical Universal”. In this case, we see how he tries to reinforce the validity of this type of Universals by intensifying the elements. Not only is it said that there is a statistic behind the Universals, but it is made explicit that this is statistically significant (something that was already explicitly understood). For this reason, and because of the excessive long-expression, it is not the best option for using this Universal type.

However, the most significant contribution, in our opinion, of this author (Dryer, 1998) on this subject is the summary he makes of the main criticisms that have been formulated over the years towards statistical Universals and tries to explain them (and argue them in some case). The different arguments Dryer offers are:

- The argument that absolute Universals are falsifiable (Dryer, 1998:4).

The first argument is that absolute Universals but not statistical Universals are falsifiable. The argument goes as follows. If one offers an absolute Universal as a hypothesis, a single exception will falsify the generalization as an absolute Universal. In this sense, absolute Universals are falsifiable. In contrast, if one offers a statistical Universal as a hypothesis, then a single exception will not falsify the generalization, nor is it clear how many exceptions one would need to falsify it.

In this case, Dryer assumes the criticism to be true. He limits himself to criticizing absolute Universals since, although they can be falsified, they cannot be tested. It will never be possible to verify 100% whether all languages ascribe to it. On the other hand, he points out that statistical Universals, which cannot be falsified according to Dryer and the critics of this type of Universals, can be tested insofar as the claims are made by allowing for an expected or unforeseen presence of languages that contradict it. However, Linguistic Typology assumes that it has not been able to determine severely and rigorously the treatment exceptions must receive to establish the differentiation between a statistical Universal and a rule that is not Universal.

- The argument that absolute Universals explain particular instances (Dryer, 1998:12).

The second argument sometimes offered in favor of absolute Universals and against statistical Universals is that absolute Universals explain instances that conform to the Universal, while statistical Universals do not. The logic of this argument is that an explanation must entail the particular facts it explains. Absolute Universals entail that each conforming language exhibits

the facts it has, while statistical Universals cannot do this. The statistical Universal that OV languages tend to be postpositional does not entail particular instances of OV languages with postpositions, since there are a few OV languages with prepositions.

To confront these claims, Dryer (1998) draws a comparison with physics and biology. He determines that working with languages is very similar to biology: one works with complex and living systems, which brings many things into play. Furthermore, he determines that the work of the linguist, if possible, is more exciting and laborious since he must determine both the general rules observed and study the exceptions observed and try to understand them. The quest for explainability is very high. To better understand this parallelism, we add a more concrete example based precisely on the anatomy of humans. A claim such as “humans are characterized by having ten fingers on their hands” is true, but it does not have to be true 100% of the time. Due to birth malformations, previous genetic problems, or subsequent accidents (as occurs in languages), this number can be modified, and we can find humans with 12 or 9 fingers, for example. This anecdote should not invalidate the possibility of proposing a rule that allows us to know better the anatomical characteristics of human beings. To expand on this concept and its relationship with biology, we recommend reading Pagel (2009).

This interest in the description and explanation of exceptions has also begun to be pursued in isolation because of the genuine interest in structures that deviate from the Universal, which is what has been called “rarum” (Iosad, 2010; Wohlgemuth & Cysouw, 2010a, 2010b). The interest in these lies in explaining the capabilities and limits of human languages (Cysouw & Wohlgemuth, 2010). Furthermore, their nature is so uncommon that a feature is usually considered rare that it does not even occur within the same family or area but in a few unrelated languages. If that *rarum* has only been identified in one language and not in similar ones, it is usually called “rarissimum” (Cysouw & Wohlgemuth, 2010). However, we find other labels that refer to this fact, such as “absolute rara” (Bickel & Nichols, 2003), “singulare rara” (Plank & Filimonova, 2000), “nonesuch rara” (Plank & Filimonova, 2000), “unicale rara” (Cysouw & Wohlgemuth, 2010) or “unique rara” (Cysouw & Wohlgemuth, 2010).

- The argument that absolute Universals tell us more about language (Dryer, 1998:13).

There is a sense in which, if we know that all languages have some property *p*, then we know more about language if than we know that most languages have property *p*. In that sense, absolute Universals tell us more about language. But this provides little argument for absolute Universals. (...) We can never even know that an absolute Universal is probably true.

Dryer (1998) is quite forceful in doubting the reliability of an absolute Universal because of the problems involved in testing it. However, what we consider more relevant about this issue

is the type of linguistic elements examined in Universals that usually result in an absolute and statistical formulation. In the first case, they are usually much more general formulations so that they can encompass any type of language and so that any language can accept the Universal as such. Therefore, the linguistic knowledge that can be obtained from such general questions is more limited than that which can be obtained from concrete and more specific formulations.

In this regard, Bickel (2010:77) comments that:

A core dimension of the ambiguity is captured by the opposition “absolute vs. statistical Universal”, although the literature uses these terms in varied ways. Many textbooks draw the boundary between absolute and statistical according to whether a sample of languages contains exceptions to a Universal. But the notion of an exception-free sample is not very revealing even if the sample contained all known languages: there is always a chance that an as yet undescribed language, or an unknown language from the past or future, will provide an exception

Such a statement reinforces the idea that absolute Universals are much more vulnerable to change, while statistical Universals are more reliable concerning their claims.

- The argument that statistical Universals may reflect differences in frequency due to historical accidents (Dryer, 1998:14).

A final type of argument that I have heard directed against statistical Universals is that the different frequencies of occurrence on which the Universal is based may simply reflect historical accident rather than a linguistic preference. If one linguistic type is more frequently attested among the set of attested languages than a contrasting type, the difference might be due to either of two sorts of explanations.

In this case, Dryer (1998) comments on different elements present when formulating a Universal statistic that ensures the independence of historical accidents. The first aspect to consider is that a sampling is taken into account (which we will develop in later sections) that guarantees a selection of languages with genetic, historical, and area independence. In addition, the functionalist motivation of this type of Universals generates that there is always a need to explain the cause of such a Universal, a fact that, according to Dryer, makes it impossible for the causes to be historical.

In the 21st century, on the other hand, we observe a remarkable change in the terminology used for references to statistical Universals. Odden (2003:38) is an example of this, as can be seen in the quotation:

The relation between absolute observational Universals and statistical tendencies of the Greenbergian type is that these are statements about observations, i.e., statements about languages and not grammar. An absolute Universal is one where 100% of human languages have a certain property, and a statistical tendency is one where a significant number, but not 100%, have the property.

This example shows the use of “statistical tendency” (and its opposite, “absolute observational Universal”), a very established form nowadays. However, these words of Odden (2003) are essential for many other reasons. On the one hand, he comments on the possibility that typological Universals can be absolute or statistical, as opposed to what some authors claim (something we will discuss in the following subsection). For example, this can be compared with the view of Kiparsky (2004), who makes explicit that UG-based Universals are all exception-free, whereas functional Universals always have. On the other hand, Odden (2003) defines these “statistical tendencies” as properties that are not in 100% of languages but in a significant number of them. In other words, we still do not find any proposal regarding establishing the maximum number of exceptions allowed to formulate a “statistical tendency”. This is something equally applicable to the proposal of Pericliev (2012:223): “linguistic tendency”, although in this case we should add the mention that the reference to statistical quality does not appear. We assume this is because it is understood that it is already included under “tendency”.

Just as in the previous paragraph, we have seen the substitution of “Universal” for “tendency” in Bickel (2007), we can observe a search for the most appropriate expression and, in this case, substitutions of the other word are attempted. That is, it seems that what does not convince Bickel (2007:2-3) is “statistical”. The result is two expressions such as “Universal preferences” and “Universal trends”. Something similar is proposed by Guevara and Scalise (2009:101), using “Universal tendencies” in their paper. In all these cases, we can see how there is an interest in not leaving the word “Universal” in oblivion, which they consider to be proper to the field and an essential part since the foundational work of Greenberg (1963), as well as the need to replace “statistical” with any other word that may remind us of the lack of application to all the languages of the world, being a rule adopted by most of them undoubtedly. This interest in discarding the term “statistical” may be due to the lack of clear statistical criteria to determine the nature of the various Universals. Words such as “preference”, “trend”, or “tendency” are much less marked than “statistical” and allow, indeed, a greater acceptance of indeterminacy when analyzing these Universals statistically.

In this same period, Moravcsik (2010:2) dares to make another proposal: “probabilistic feature”, as mentioned in the quotation:

All language Universals are merely hypotheses. Although they may hold exceptionless or with a certain degree of probability for a given sample of languages, there is no assurance that additional languages will not turn an absolute Universal into a probabilistic feature, or a probabilistic feature into one that is merely possible but not likely. The only kind of cross-linguistic statement that is impervious to refutation is the weakest kind that simply says that a structural pattern is possible in human languages. This is so because once a pattern has been identified in a language, it must of course be possible.

Another author who opts for the maintenance of the statistical part and a replacement of the Universal part is Siemund (2011:11), with the term: “statistical generalization”.

This nomenclature is not perfect. In the first place, it offers a synonym for “statistic”, such as “probabilistic”, which goes against the other authors’ tendency. No probabilistic criterion is offered to classify Universals, but such a term is used. On the other hand, the word “Universal”, globally accepted and with tradition in Typology, is replaced by a less proper option as a feature. From this quotation, we can highlight other issues that we consider to be very accurate. In addition to recalling the nature of the hypotheses we have mentioned above, it reflects pretty clearly the limitations of determining which rules are valid or which are not. Finally, it highlights the type of claims that are more valid from a statistical acceptance point of view; however, these, being so general, are not of interest to Linguistic Typology to understand the behavior of languages.

Moravcsik (2012:243), years later, prefers to work with the expression “cross-linguistic preferences”. The former word both clarifies and delimits the typological nature of claims with exceptions. In our opinion, the term preference somewhat minimizes the power of the restrictions behind the various Universals.

In Kirkegaard (2016:1-3), we note an essential terminological variation: “language patterns”, “Universal generalization”, “statistical pattern”, and “probabilistic tendencies”. In all cases, reference is made to the “statistical Universal” concept. Such variety can confuse the reader, as the same concept is always referred to in a rather unsubstantial way. Many of these labels are redundant or even fail to show the information to which they should refer and which is perfectly captured in the expressions highlighted above.

We find a proposal for fixing the criteria for the differentiation between “absolute Universals”, “statistical Universals”, and “non-Universals” in Kirkegaard (2016:3). According to him, absolute Universals are those rules whose result is 100%. Statistical Universals are those rules whose result is between 50% and 99.99%. Therefore, rules with a result less than 50% will be considered as non-Universals. We openly disagree with the proposal of Kirkegaard (2016). We do not think it is reasonable for a rule present in 50.1% of the languages of the world (or the sample) to be considered statistically Universal and, therefore, valuable for study in Linguistic Typology. Such a result does not show any kind of clear preference for one direction. In a case in which one had to predict the behavior of a linguistic element in a new language, the probability of getting it right or wrong would be practically the same since the information is not representative of any preference.

We leave aside the treatment of other labels that we consider to be very wrong, such as “eu-roversal” (Loporcaro, 2011:55), a Universal that, above all, has a high degree of acceptance in the European area, as we consider them to be out of place and misleading, as they have no basis in any kind of linguistic criteria.

Finally, in recent years, there have been few new proposals. Most focus on leveraging previously existing expressions, such as the distinction between Universals and near-Universals in Comrie (2020). Guzmán Naranjo and Becker (2022) prefer the usage seen earlier in Miestamo et al. (2016): “crosslinguistic tendencies”.

Grandi (2020:74), instead, focuses the distinction between “Universali” (Universals) and “tendenze” (tendencies). In such a decision we see how he wants to simplify as much as possible and leaves the original term only for the cases without exceptions and the tendencies to indicate the cases in which some exception appears.

Finally, we would like to highlight the contribution of Sansò (2022). In this paper, we observe a different application of the terms, which can lead to confusion. For example, one can read how: “l’individuazione delle caratteristiche presenti in tutte le lingue, o almeno nella stragrande maggioranza di esse (Universali assoluti)” (Sansò, 2022:55)¹. This quotation comments on how the formulation of rules present in all languages, or at least in a significant part, is known as “absolute Universals”. That is, both situations are being merged under the same “absolute Universal” label. As can be understood, such a decision does not seem very wise because of the problems it may entail.

In fact, Sansò (2022:55), when he needs to establish the distinction between absolute and statistical Universals, resorts to the expressions “Universali assoluti propriamente detti”². and “Universali assoluti statistici/probabilistici”³. Such expressions do not seem very productive or transparent since, in our opinion, they further complicate a distinction that was already problematic because of the facts discussed in this section.

Finally, from Sansò (2022:55), we would like to highlight the quote:

In tipologia si parla di Universali assoluti per indicare caratteristiche che tutte le lingue possiedono senza eccezioni, o, in un’accezione un po’ più ampia, per designare caratteristiche che la stragrande maggioranza delle lingue possiede. Un termine più preciso per definire gli Universali assoluti e per distinguerli da quelli implicazionali, ma che non ha una tradizione nella terminologia italiana, è Universali senza restrizioni (traduzioni dell’inglese unrestricted),

¹This could be translated into English as: “the identification of features present in all languages, or at least in the vast majority of them (absolute Universals)”

²This could be translated into English as: “actual absolute Universals”

³This could be translated into English as: “statistical/probabilistic absolute Universals”

a loro volta distinti in assoluti (“in tutte le lingue X”) e statistici o probabilistici (“nella stragrande maggioranza delle lingue, X”)⁴.

In this case, the Italian linguist puts absolute and implicational Universals on the same level. In general, there is a great deal of mixing of the different labels referring to different modalities, which can generate confusion. However, this quote allows us to summarize the three key points related to this subsection very well.

First, there is a wide variety of labels to refer, in either case, to the presence or absence of exceptions in the formulation of Universals. This vast terminology is no more than the consequence of the lack of resolution of the naturalization of exceptions in Linguistics. We believe that the fact that the historical label is “Universal” may condition more reluctance to believe that a rule with that name can have exceptions. However, experts in the field should be able to overcome such a stumbling block since there is no need for such a correspondence between this term and its use in a given discipline.

Second, the problem with the number of exceptions that can be tolerated for the Universal to remain valid has not yet been solved. The few attempts that have been observed have not been satisfactory for typologists.

Third, the labels proper to the distinction of the presence or absence of exceptions are sometimes mixed with other labels referring to other modalities, implying that they are often synonymous or, at least, that they always go hand in hand. This is the aspect that we wanted to discuss in the following section.

6.4 Labels that Are Used for Different Variables

We believe that the most organized and respectful way to classify Universals is by showing in isolation the different variables about the different classifications and labels of Universals. This is somewhat similar to what Moravcsik (2010:1) does, although in her case, she only does it around two variables and not three as we do:

Like all generalizations, language Universal statements may vary in modality and domain. Concerning modality, some Universals are stated as exceptionless, holding for every member

⁴This could be translated into English as: In Typology we speak of absolute Universals to designate characteristics that all languages possess without exception, or, in a somewhat broader sense, to designate characteristics that the vast majority of languages possess. A more precise term to define absolute Universals and to distinguish them from implicational Universals, but which has no tradition in Italian terminology, is “Universali senza restrizioni” (the translation from English “unrestricted”), which are in turn distinguished into absolute (“in all languages X”) and statistical or probabilistic (“in the vast majority of languages, X”).

of their universe. These are called absolute Universals. Others are probabilistic (also called statistical), stated as holding for most but not all languages. Concerning domain, Universals may be stated for the entire universe of languages or a contextually delimited sub-universe of them. The former are unrestricted Universals, and the latter are restricted Universals (also labeled implicational or typological Universals).

This same view of the topic is held by Comrie (1989:19). This linguist is the one who offers the best representative examples of the different combinations in the following quote:

There are absolute non-implicational Universals, such as all languages have vowels. There are absolute implicational Universals, such as if a language has first/second person reflexives, then it has third person reflexives. There are non-implicational tendencies, such as nearly all languages having nasal consonants (although some Salishan languages have no nasal consonants). Finally, there are implicational tendencies, such as: if a language has SOV basic word order, it will probably have postpositions (but Persian, for instance, is SOV with prepositions).

As can be seen, the distinction between surface and deep Universals does not come into play in these combinations. This is mainly due to the lack of flexibility of the deep Universals, which do not present either examples in the form of implication or with exceptions, given their nature.

However, once we have shown all the variables aseptically, we should remember that the possible combinations resulting from mixing the options of the different variables are not all observed with the same frequency. There is a tendency to observe different combinations, the most frequent and exciting being those corresponding to superficial, implicational, and statistical Universals (and their antonyms: deep, unrestricted, and absolute Universals). This is why many authors, either because they want to generalize, because of confusion, or because they want to emphasize this close relationship, interchange or mix some labels.

Greenberg et al. (1963:259) even uses new labels to reflect such relations, such as “statistical correlations”. This is the product of assuming the implicational Universals as inherent with the presence of exceptions. This same concept is called by Hawkins (1980:194) “statistical implicational Universals”.

Steedman (2009) and Grandi (2020) are examples of authors who consider “implicational” as a synonym for “surface”. They believe that Universals produced from functionalism are always in conditional format. They are not wrong in considering that such Universals are the most relevant and representative.

Bickel (2010), on the other hand, focuses on emphasizing that the surface-deep opposition is equivalent to absolute-statistical, showing the similarities. Again, such assimilation (although reductionist) is accurate, as the research methodology of surface Universals has been shown to

favor the presence of exceptions. In contrast, deep Universals contradict the presence of exceptions in generativist logic.

Finally, Grandi (2020:62-63) comments that: “Gli Universali assoluti sanciscono la presenza (o l’assenza) di una particolare proprietà in ogni lingua storico-naturale, senza fare riferimento ad alcun altro parametro e senza stabilire correlazioni fra tratti differenti”⁵. That is, he defines absolute Universals as the characterization of a property without reference to any other parameter. In other words, he is using a definition that would fit perfectly for the term “unrestricted Universal” and not “absolute Universal”.

Apart from the various evident marks found in the literature, this same point of view is perceived in many other publications. However, it has not been as explicit as in the examples shown. A notable number of expert authors blur the boundaries between the abovementioned terms. We do not deny that they are correct because, as we have seen in the various definitions and examples, there are obvious synergies between particular configurations of Universal, conditioned by the research methodology, which, at the same time, is conditioned by what the researcher understands by Universal.

However, despite what we have seen when dealing in a detailed and deep way with Universals in language, we believe it is inaccurate to generalize that, for example, Greenberg’s Universals are implicational and statistical. As will be seen in this thesis, some of his Universals, although less known, do not have an implicational structure and apply to 100% of languages, something that can sometimes be forgotten.

In conclusion, despite the overlaps we have seen, we can classify the labels seen in a summarised way with their most prototypical ascription in terms of their meaning as in tables 6.1, 6.2, and 6.3. However, we should not forget what we have seen in this last section, as the labels applied to the presence of exceptions are closely related to the labels used to conceive the Universal and are often used as synonyms.

⁵This could be translated into English as: Absolute Universals sanction the presence (or absence) of a particular property in every natural-historical language, without reference to any other parameter and without establishing correlations between different traits.

Labels to Distinguish the Conception of Universal			
Surface	Language	Property	Features of Language
Deep	Features	Observational	i-Universals
Substantive	Aristotelian	Architectural	e-Universals
Definitional	Galilean	Cross-linguistic	Probabilistic
Formal	Chomskyan	Typological Generalisation	True
Prediction	Greenbergian	Grammatical Features	Language Patterns
Linguistic			

Table 6.1: Summary of Labels Regarding the Conception of Universals

Labels to Distinguish the Structure of the Universal	
Unconditional	Restricted Equivalence
Conditional	Biconditional
Implicational	Typological Correlation
Non-implicational	Typological Implication
Correlations	Implicational Hierarchy
Restricted	Unrestricted

Table 6.2: Summary of Labels Regarding the Structure of Universals

Labels to Distinguish the Presence of Exceptions		
Statistical	Significant Statistical	Cross-linguistic Preferences
Absolute	Preferences	Universal Generalization
Non-universal	Trends	Statistical Pattern
Near	Cross-linguistic Generalization	Probabilistic Pattern
Tendencies	Cross-linguistic Patterns	Cross-linguistic Tendencies
Nearly Exceptionless	Statistical Generalization	Almost Trivial Universals
Exceptionless	Linguistic Tendency	

Table 6.3: Summary of Labels Regarding the Exceptions of Universals

7

Methodologies for Universals Research

Contents

7.1	Universals Formulated from the Armchair	66
7.2	Universals Formulated from the Laboratory	73
7.3	Universals Formulated from the Field	75

This chapter reviews the main existing methods for conducting Universals research. There have been different avenues of research based on different typologies of data sources. We put these studies in order and show the main characteristics of these studies, as well as the fundamental benefits and limitations. Showing the different possibilities helps us to understand better the differences between the different existing conceptions of Universal or even the evolution of the conception of Universal within the same approach.

To classify these origins, we have found it very interesting to adapt a proposal made by H. Clark and Bangerter (2004) that they use to deal with reference in pragmatics. They offer a division into three parts, which they name according to the place where it is produced:

1. Universals formulated from the armchair
2. Universals formulated from the laboratory
3. Universals formulated from the field

This classification and nomenclature that they propose for their pragmatic study, we believe, is valid for any research in Linguistics (even in other humanities disciplines, for example, in anthropology) and, therefore, we believe that we can offer an adaptation to understand better the provenance of the data with which Universals are formulated.

A good summary of the three possibilities we present would be as follows:

- **Armchair.** The researcher uses introspection to support the claims. The examples are from languages known to the researcher and are selected. Another typology of studies based on comparing information contained in grammars of different authors can also be included in this group. The first case uses the most proper armchair in Chomskyan Universals (armchair), and the second uses the most proper armchair in Greenbergian Universals, that is, consulting grammars (library armchair).
- **Laboratory.** It corresponds to the study of language based on the controlled productions of different subjects obtained by the linguist. A study is conducted with participants who respond to questions related to the linguistic study.
- **Field.** It corresponds to the study of language based on real, uncontrolled, and unconditioned textual productions. The linguist processes such data to study the linguistic phenomenon that concerns the researcher.

7.1 Universals Formulated from the Armchair

Most of the time, we find examples of a methodology based on intuition in armchair Linguistics. The linguist is considered an expert on the subject and can generate examples to support the theory. Alternatively, in other words, the examples with which the study's author generates the linguist works.

This means that the data available to support the theory are pretty limited and do not enjoy 100% independence since there can always be suspicions about the appropriateness of the examples. It is also easier for counterexamples not foreseen by the author to appear, who tends to simplify the phenomenon to the offered representation. However, given the simplicity of this methodology, it is the ideal way to begin studying a given phenomenon for which no data are available.

This methodology is very well related to what we have presented above as deep Universals. That is, Universals within the logic of Universal Grammar arise from approaches such as the one described in this section. Since the languages analyzed are usually mastered by the linguist studying the phenomenon, this allows for a much more detailed (or deep) approach to the linguistic element analyzed, which is very much in line with generativism. The apparent limits cause it to not be possible to analyze, at the same time, a large number of languages, as is the case with other approaches.

Chomsky (1965:18) himself stresses the value of this method as the best method for hypothesizing:

The actual data of linguistic performance will provide much evidence for determining the concreteness of hypotheses about underlying linguistic structure, along with introspective reports (by the native speaker, or the linguist who has learned the language).

Moreover, it seems that Chomsky (1965:20) is unwilling to disregard the advantages of other approaches, so he chooses to give importance to the methodology used to obtain these data for the formulation of Universals:

One may ask whether the necessity for present-day Linguistics to give such priority to introspective evidence and to the linguistic intuition of the native speaker excludes it from the domain of science. [...] In Linguistics, it seems to me that sharpening of the data by more objective tests is a matter of small importance for the problems at hand.

Such a view can be confirmed by Gilquin and Gries (2009:2), who, with a broader perspective, assert that:

As is well known, a fundamental change of perspective was introduced with the rise of formal syntax, or more specifically, transformational-generative grammar, in the 1950s and 1960s. This went hand in hand with abandoning collected data of the above kind and introducing very informally collected linguistic acceptability judgments (largely by the analyst him/herself) as the primary source of data. Paradoxically, the lack of sophistication adopted in gathering the judgments was considered convenient and a virtue.

For a critique of this methodology, we propose the extension in Tanaka-Ishii (2021) and Tomasello (1999, 2003).

The same view is held by scholars of the Universals in pragmatics, such as Searle (1976, 1994), Kryk (1990) or Austin (1962), who make explicit their introspective approach to the phenomenon, but the possibility of testing such claims using more concrete plans such as the two methods appears.

In the foundational publication of Universals, we find some authors like Ullmann (1963) who carries out this approach used and described by Chomsky. In this chapter, he begins by philosophizing about what the researcher believes to be Universal in semantics. To do so, he illustrates his argumentation with different examples from different languages. However, we are not dealing with a systematic study of the world’s languages but with a first approach to a phenomenon that, later, Linguistic Typology will have to confirm or certify through conclusions drawn from other data. It is also important to mention that Ullman’s proposal is not based on the examples he offers, but that they serve, a posteriori, for a more concrete exemplification of the linguist’s intuition about the subject he is studying (Haddington et al., 2005).

It is important to remember that this debate between a more rational or empirical approach to linguistic phenomena is not unique to the study of Universals (Mairal & Gil, 2004). It is a classic dilemma in Linguistics (and in other sciences). However, we also observed this debate about Universals directly and explicitly between two researchers. In short, Itkonen (1980a, 1980b) argues that Linguistics is more like the rational sciences than the empirical sciences. He summarizes the conflict nicely: “Either the grammarian investigates sentences invented by himself or he investigates a corpus of actual utterances. (...) This conflict of opinions can be summed up as a conflict between intuition and observation” (Itkonen, 1980b:335). In this case, Itkonen argues there is no added value in researching sentences from a corpus rather than invented ones. He also takes the opportunity to mention that many of the linguists who advocate quantitative corpus research then fail to practice it as an argument to defend his position. In contrast, Dahl (1980) directly criticizes Itkonen. The researcher points out the weaknesses (already seen) of deep Universals and the incompatibility of an armchair approach to formulating cross-linguistic Universals.

Although this methodology, based on intuition, can also be used in typological studies on Universals (especially in the early days of the discipline), the use of data extracted from grammars, known precisely as “grammar-based Typology”, was and is more widely accepted (Schnell & Schiborr, 2022:172). In this case, we decided to include such a source also in Linguistics made from the armchair, although if we want to be more precise and continue with the metaphor, it would not be the armchair of a house but the armchair of a library. This happens because the linguist bases his claims on second-hand data, as also indicated by Mauri (2022b:51): “ogni volta che si consulta una grammatica descrittiva si fa affidamento sul giudizio e sulle intuizioni di qualcun altro”¹. The grammars of the different languages consulted come from different linguists, and we find an important mix of sources (*in situ* documentation, intuition, etc.). This point of view and

¹This could be translated into English as: “every time you consult a descriptive grammar you rely on someone else’s judgment and intuition”.

classification that we express is supported by Daniel (2010), who uses, on the other hand, the term “typological armchair”. To expand on this aspect, we recommend reading Mauri (2022b).

Another reason why we include such studies in this subsection is because of the similarity of benefits and limitations obtained with the comparison between grammars if we compare them with the case presented above. Again, we start from general knowledge (not *ad hoc* for the phenomenon) from which some example is rescued, which allows us to know how such a language behaves. All this is filtered by the linguist’s perspective in charge of such grammar.

The only difference we find is that, in this case, it is possible to analyze many languages. Also, the data obtained are much more general, usually knowing only the most prototypical structures and without having access to much more borderline or atypical occurrences, which can emerge in studies where the linguist is an expert in the language being analyzed.

In fact, Croft (2003:29) asserts that: “The data sources used by typologists are: native speaker elicitation, texts, and descriptive grammars”. Thus, there is a coincidence between our proposal and Croft’s, in which it is understood that Typology can be studied from the armchair only with grammar (library armchair), which evidences the incompatibility between intuition and Typology. In short, Croft (2003:49) assigns the empirical method to Typology and the rationalist to the generativist:

One of the features that distinguishes the typological method of discovering constraints on possible language types is the empirical method applied to the problem. Suppose a typologist wants to find restrictions on possible relative clause structures, for example. In that case, he or she gathers a large sample of languages and observes which possible relative clause types are present and which are absent. The attested language types motivate the restrictions on logically possible language types. If there is a gap in the attested language types, then it is provisionally assumed that the gap represents a constraint on a possible language, and explanations are sought for the gap. This is the inductive method, which must be used in constructing generalizations from empirical data. In contrast, the generative approach uses a rationalist deductive method, in which it is argued that certain analyses of a single language represent Universals of human language because they cannot possibly be learned by a child (the ‘poverty of the stimulus’ argument).

Even a clear defender of the use of corpora for typological research, such as Levshina (2022:2), sees several advantages in the use of grammars for quite specific and determinate aspects of formulating Universals:

It seems that no corpora at present can replace the traditional type of typological data based on language description in reference grammars at least, not for all research questions. Some fundamental linguistic properties, e.g. the presence of a definite article or the number of cases in a language, are much easier to find in a grammar than to extract from a corpus. The practicality of using grammars, however, comes at the cost of having to rely on someone else’s judgements and intuitions based on their own more or less representative corpora. Using corpora from the very beginning will allow us to investigate language use more directly and

in greater detail (although corpora, of course, also represent the result of someone's choices, from sampling to word segmentation and annotation).

In this case, we slightly disagree with such a quote. As will be seen in this thesis, it is equally easy to check with Universal corpora the presence of a given feature in languages or the number of cases. We believe that even the study using corpora is more effective since it can help to see elements considered borderline in grammar but existing in languages; in addition, the possibility of having a high degree of automation makes such a search faster, more effective, and less costly.

When using grammars, it is possible to speak of a method for formulating Universals with them. In this case, the reference by Mattioli (2020), who starts from the premise that there is no ideal grammar to work with the data of such a language, seems to us to be very accurate. The criteria he points out for the choice of grammars (and the discarding of others) are the following:

- The most recent one.
- The most extensive one.
- The most accessible one.

The conjunction of the different elements previously described in Mattioli (2020) allows us to discriminate which grammar is the most appropriate for a study of this type. The more recent version, *a priori*, allows us to obtain corrections of data that might have been previously stated erroneously. The more extensive grammar, presumably, guarantees a higher level of detail of the phenomena described. The fact that this grammar is easy to obtain simply favors the work and effectiveness of the researcher, who would have to devote much time to obtaining other grammars, multiplied by the number of languages to be analyzed. To know the available grammars of a language, tools such as Glottolog (Hammarström, Forkel, Haspelmath, & Bank, 2023) are essential. In such a resource, the sources associated with each language appear very orderly and are helpful.

One of the classic and logical problems when working with different grammar of the world's languages is the disparity in the linguistic analysis of the languages. A particular view of Linguistics may influence such analysis or contain terminology that does not allow comparison with other languages (or may even be confusing). The potential lack of comparability of grammars by different authors has been an issue that has enjoyed interest in the field and where a pretty adequate solution has been proposed (Mauri, 2022b). Comrie and Smith (1977) decided to make a questionnaire for Typology to be used in linguistic documentation. This has been used to create elementary

grammars for easier comparison (although there may always be small details that differ). This has given rise to linguistic series edited by North Holland, Croom Helm, and Routledge, which have specialized in this publication of homogenized grammars. The main problem is the limited number of grammars in these series since this number is much lower than the documented languages.

Mauri (2022b:48-49) perfectly points out this evolution in grammars throughout history:

Nell'ultimo quarto di secolo c'è stata una crescente convergenza nei modi in cui le lingue sono descritte. Se analizziamo le grammatiche scritte nel periodo dal 1965 al 1975 circa, si trova una notevole variazione nelle categorie descrittive impiegate, alcune più legate alla grammatica tradizionale, altre ad approcci strutturalisti (come la tagmemica), altre ancora inserite nel quadro generativista. Dopo il 1975 circa questa varietà diminuisce per una concorrenza di fattori: il graduale declino dell'uso della grammatica generativa come mezzo per descrivere intere lingue si accompagna a un'uscita di moda degli approcci strutturaliste, e nel frattempo si osserva un significativo aumento di ricerche condotte nell'ambito della tipologia linguistica, i cui risultati sono stati di grande utilità per i linguisti descrittivi².

The above quote explains how, between 1965 and 1975, there was a significant variation in the descriptive categories used in grammar, some linked to a more traditional approach, others more structuralist, and others more generativist. In 1975, according to Mauri, there was a change of trend, with less use of generative grammar to describe languages and an increased interest in Linguistic Typology and more homogeneous and less marked methods of linguistic description.

We would also like to add a special mention to databases such as the World Atlas of Languages Structures (WALS) (Dryer & Haspelmath, 2013). In these cases, we are faced with a digitized database, with the unification of data from multiple languages worldwide. The conveniences offered are obvious: ease of access to data, effective visualization, etc. However, we must not forget that the source of these databases is information from different grammars of the world. Therefore, we are simply dealing with an updated, more convenient version of the grammar approach. This means that the limitations of the grammar approach are being carried over. However, the user may think that this new type of tool is an innovation in data or methodology.

To begin with, we continue to assign the different languages to a single linguistic type without considering the possibility of expressing different nuances (such as whether it is rigid or presents occurrences in another sense). Moreover, this linguistic type is assigned thanks to the mention of some linguists in some grammar, without any confirmation or assurance of the correctness of this

²This could be translated into English as: "In the last quarter century, there has been an increasing convergence in the ways languages are described. If we analyse the grammars written in the period from 1965 to 1975 or so, we find a considerable variation in the descriptive categories employed, some more related to traditional grammar, others to structuralist approaches (such as tagmemics), and still others within the generativist framework. After 1975 or so this variety diminishes due to a competition of factors: the gradual decline in the use of generativist grammar as a means of describing whole languages is accompanied by a fall out of fashion of structuralist approaches, and in the meantime there is a significant increase in research conducted in the field of Linguistic Typology, the results of which have been of great use to descriptive linguists".

data.

It is also important to remember that, in this case, the list of characteristics for which we have data is also very limited. These, moreover, are the most canonical or representative of a language without having more specific data on borderline structures or those not previously treated by linguists (all the information is already known). In other words, all the available information is information that other linguists have already studied and appears there.

The data come from different grammars without considering any kind of unifying criteria. This may result in comparing data from different languages that are not comparable or have been understood differently.

Of course, having only one category that summarizes a language's linguistic type means that metrics cannot be obtained beyond the information provided. Different examples are unavailable, and no review is based on different contexts.

All these complaints can also be applied to a new database recently published, equivalent to WALS, called Grambank (Skirgård et al., 2023). In this case, we face the same idea, and the errors or limitations are perpetuated. One of the differences announced by the creators, which is quite interesting, is the possibility of avoiding the assignment of a label in terms of dominance, as it happens in WALS, and allowing the activation of two labels at the same time if the language contains occurrences in two senses, each type having its entry. This makes it possible to escape from the concept of dominance alone and also to have information on borderline structures. However, after a few simple checks, we see this is not the case, leading to further errors. For example, regarding the direction of the adposition, WALS shows the dominance label, which in this case would be “preposition” in English. In Grambank, we expect to find the value “1” (meaning the presence of the element, regardless of the quantity) in both the preposition and postposition entries. This is because we find English postpositions such as “ago”. However, in the postposition entry, we find the value 0. Only thanks to a field approach, as we will see, we can obtain this information in a fast, homogeneous and accurate way.

In summary, we have seen the possibility of formulating Universals either from one's intuition with the creation of examples to support the linguist's claims or from comparative data extracted from different grammars of different languages. The first example is the most suitable for generativist studies, and the second for typological studies. Despite the various limitations discussed above, it is an excellent starting point for approaching the different phenomena of the world's languages.

7.2 Universals Formulated from the Laboratory

Another possibility to obtain the linguistic information necessary for research is the design of an experiment. The objective is to obtain data specifically sought in certain subjects in an elicited manner. To do this, a questionnaire must be created to elicit the object of study. In other words, to create instructions or questions that make the subject produce what the researcher is interested in. Secondly, the subjects who will participate in the study must be selected. This must be representative from a linguistic point of view to allow the claims to have the most negligible bias from a given group (if the study is general). The sample size must also be sufficiently representative.

Such studies are very successful in psycholinguistics, for example. Already in the foundational publication in Osgood (1963), the use of tests for obtaining psycholinguistic data is mentioned. We believe their use is exciting, especially when the linguist already knows the phenomenon under study (thanks to the previous methodology or other studies on the subject). Thanks to the elicited use, this verification in different confirmed speakers allows us to obtain multiple sources simultaneously, showing a much more precise and accurate tendency than relying on a single source. In addition, synonymous, non-canonical structures may appear, which the researcher may not have anticipated. The analyzed speakers can validate and weigh them through their use in such a study.

In addition to the benefit outlined above, this approach allows new metrics to be obtained. It is possible to monitor speaker response, measure response times for comparison, and extract patterns related to a given linguistic stimulus or structure. In other words, applied to the case of Universals, it can help to strengthen various functional theories based on ease of processing, for example.

However, this approach also offers some limitations that should be considered. Creating a balanced and representative sample across all terms in multiple languages is tricky. In many cases, an evident bias is often overlooked because of the impossibility of providing better data (even in very accessible and European languages). The challenge of obtaining a group that can carry out the proposed study is much more significant in the case of much less recognized languages.

Moreover, the lack of means available to the linguist often does not allow the development of “laboratory” studies in a given language with large and representative samples. The means to repeat such an application in hundreds of languages, many of them thousands of kilometers away, is almost unimaginable. The means required are beyond the means of most linguists.

Finally, we must emphasize that the data produced in the laboratory are not attractive if one seeks a quantitative measurement of the occurrence of the phenomenon under normal conditions. The linguistic element under analysis is elicited (Siemund, 2011:11). This means that the output produced by the study subjects is not free and is conditioned both by the previous requirements of the linguist and by the mere fact that they are answering a test of these characteristics.

This methodology can also be applied when documenting a language, as shown by Pym (1979). Such a selection can be made, and a questionnaire can be prepared to obtain grammatical data in the studied language so that grammar can be elaborated. Subsequently, this information can be used for Typology and Universals. We already find proposals in this sense in the typological research of the 19th century, as in von der Gabelentz (1869), intended to reduce bias.

It is important to mention that not all questionnaires carried out in this sense automatically correspond to a laboratory approach. The most frequent in linguistic documentation, especially in recent years, is using questionnaires only to homogenize the labels used and the linguistic theories put forward. In other words, the documenter has to fill in this questionnaire based on the observation of the speakers of the language being analyzed. This means the study is in the field, although in a more outstanding order than without using such questionnaires. Examples are the documentations of Kobon or Imbabura languages for the Essential Grammars series for North Holland (Cole, 1982; Davies, 1981), Japanese or Kannada for Routledge (Hinds, 1988; Sridhar, 1990), or Tamil and Bubungo for Croom Helm (Asher, 1985; Schaub, 1985). The guidelines followed for such homogenization are already found in Comrie and Smith (1977), as we mentioned. Subsequently, we find projects with this same interest, such as Eurotyp (European Science Foundation, 1993).

The proposal of Max Planck Institute for Evolutionary Anthropology (2015) is somewhat different since, in this case, they present the “stimulus kits”. In this case, we do observe the will to select a representative sample of speakers, isolate it from its context, and elicit responses from them, as applied, for example, by Burton and Matthewson (2015).

We do not find many approaches from the laboratory to study Universals, at least in the case of the most classical and well-known Universals. This methodology seems more suitable for typological studies of Universals in semantics or pragmatics (Kasper, 1992; Kasper & Blum-Kulka, 1993; Rose, 1992). We recommend extending the various examples of tasks that can be carried out in a publication in pragmatics (Kasper & Dahl, 1991). If other methodologies provide the same information, they are preferred to this method. It is reserved for cases with limitations in other possibilities in pragmatics.

In the case of intonation studies of world languages, the use of the laboratory also seems sensible. However, researchers in the field point out that such studies are not very productive (Hirst & Di Cristo, 1998). They do not show completely reliable data, and it is not easy to achieve the means to carry out a decent representativeness.

This view is also shared by Croft (2003:29), who states that:

Reliance on native speakers is unfortunately impractical if a large and diverse sample is desired; it also has difficulties. The process of eliciting grammatical information is not a natural language situation, so the data provided do not always accurately represent actual language use. In particular, the native language consulted does not report actual usage but instead, an unsystematic perception of usage, colored by social attitudes toward the speech form and even the interrogator.

In other words, therefore, the difficulties of conducting such a study and the limitations of the data obtained are undesirable. The use of language is not free of bias, and the cost is very high. For this reason, there is currently an increased interest in the field methodology.

7.3 Universals Formulated from the Field

In this case, the linguistic data are obtained directly and naturally. The productions are considered natural because they appear in a free (unconditioned) way and are simply collected by the researcher without any kind of direct intervention. The linguist's role includes analyzing the actual occurrences captured, whatever they may be.

In this case, obtaining new metrics not foreseen in the two previous methodologies is possible. We refer to the weight of a particular phenomenon or structure as opposed to others (objects preceding the verb, for example) in a language. In the armchair case, such quantification is impossible since we deal with isolated examples. In the case of the laboratory, since it does not have spontaneous and natural productions, being influenced by the linguist, it does not offer reliable data either. The linguist seeks the occurrences thanks to the approach of specific situations.

Another benefit of such a proposal is the possibility of obtaining more borderline or less representative structures that may occur in everyday language and that may not appear in grammars or a test designed by a linguist, given the space and time limits of a questionnaire. The amount of data available is potentially much more prominent. Language experts must also label and process the information captured in the speakers' occurrences. However, in this case, the interpretative weight of the person in charge is much less, favoring a greater homogenization for comparison by the linguist in charge. This fact can also be read in the words of Levshina (2022:7):

Reference grammars are usually good for understanding the main strategy of expressing a certain meaning, but they can be less useful for describing less frequent, context-dependent ones. Moreover, this descriptive information does not reflect the gradient nature of language use (Levshina 2019; Ponti et al. 2019). This leads to data reduction and bias towards investigating categorical features with a bimodal cross-linguistic distribution (Walchli 2009).

There is also the risk of being unable to formulate new theories or generate new knowledge if grammars are always used to formulate Universals. The elements of interest to the linguist responsible for such a publication may not be the same as those of the current researcher, which means that not all the data are complete. This is something that is perfectly explained by Daniel (2010:14):

some of the blame for typologists mistakes and misinterpretations has to be laid on. The latter vary not only in quality and reliability but also in grain. Even a reliable and detailed grammar not provide necessary information simply because an issue of interest might not have been recognized as such at the time when the grammar was written. An example of this is the volumes of the Handbook of American Indian Languages (Boas 1911-22). While these are very thorough descriptions, they prove to be of little help in answering many questions typologists started to ask years later. The best data for non-first-hand analysis are indisputably texts. These are closest to actual language use and as theory-free a type of data as possible (more so for morphology and syntax than for phonetics and phonology). Much effort has been put recently into improving practices of language documentation, including online representation (graphic, acoustic, and later also visual).

It is also necessary to mention the advantages of such an approach from a materialist point of view, especially in recent years. Just as it has been challenging to have a large number of grammars (depending on the location of the researcher) and to be able to carry out questionnaires in multiple locations around the world (for most researchers), homogeneously annotated data from texts in different languages of the world are often available openly online. This trend, in our opinion, favors the liberation of science, which makes it much more accessible to everyone (if one has an internet connection) and, therefore, more equitable.

Initially, typological studies that wanted to work with this method did so with coexistence in the different places of interest for the linguist. The linguist collects and notes everything, as is done, for example, by Burquest (1973). In recent years, on the other hand, the preference has been to start from already annotated corpora (for this or other purposes) and use them to extract new knowledge about the functioning of languages.

Therefore, in recent years, this type of approach is often synonymous with the use of digitized corpora. To extend the fundamental concepts of Corpus Linguistics for application to token-based Typology, we recommend Barth and Schnell (2022). These are the ones that seem to be most used in Typology to carry out inductive research. This trend is quite evident, especially in morphological or syntactic questions. It is more complicated to work on semantics or pragmatics, for example,

through corpora, due to the limitations that exist in current annotation systems, and that causes a much smaller number of available databases to work with. We now note a growing interest in this aspect by seeing different publications proposing corpus annotation databases on specific aspects related to pragmatics or semantics (Allen & Core, 1997; Core & Schubert, 1996; Leech & Weisser, 2003; Weisser, 2003).

Some authors go further and do not consider the use of corpus only as a “set of methods” but consider it as a theory of its own (Gilquin & Gries, 2009:5). It is also important to note that the mere appearance of a corpus in an investigation does not guarantee that it is a field investigation. Levshina (2019) conveniently points out the difference between a “corpus-driven” and a “corpus-illustrated” study. In the former, we find a willingness to draw conclusions or generate scientific hypotheses thanks to the data of a given corpus. In the second case, only a linguistic corpus provides an example to support the previously formulated theory. In this case, the corpus is not the source of the research. It only serves as a source of examples, substituting the one the researcher could give from his armchair.

The corpus study’s characteristics are not detailed in this section. The main reason is that it is the methodology used in this doctoral thesis, and, therefore, sections are devoted to it later to discuss in depth different characteristics that influence the study. Although we would like to recommend the publication of Croft (2003) to expand on the existing limitations of a corpus-based study methodology (and also for grammars). We also recommend Futrell et al. (2020) for the weaknesses that may make corpora unreliable for making claims about grammars.

Thus, in short, we believe that the most sensible thing to do when one wants to study in depth any typological aspect (extensible to other linguistic questions) is, precisely, to combine the different methodologies or sources presented here. This fact will allow a much richer reading of the phenomenon. Such a combination allows us to mitigate the biases we have mentioned of some methods among them and confirms the observed tendency. We already see this in proposals for other linguistic research such as that of Masini and Mattiola (2019:282). In such a case, lists and what they call an “indagine multi-livello” are investigated. That is, a multi-level investigation is proposed. These levels are labeled as horizontal, intermediate, and vertical. The horizontal level would refer to a large-scale typological study using grammars and a representative set of 250-300 languages. The intermediate level would refer to an analysis of isolated texts in 20-50 languages to check whether such data from grammar are embodied in real productions unrelated to such previous sources. The vertical one would refer to a study with large corpora in a set of 2-5 languages of the previously analyzed facts.

UNIVERSITAT ROVIRA I VIRGILI

GLU - A GRADIENT MODEL FOR LANGUAGE UNIVERSALS. A COMPUTATIONAL REVISITATION OF GREENBERG UNIVERSALS

Antoni Brosa Rodríguez

8

Universals & Linguistic Level

Contents

8.1	Syntactic Universals	80
8.2	Morphological Universals	81
8.3	Semantic Universals	82
8.4	Pragmatic Universals	83
8.5	Phonological Universals	84

This chapter briefly shows the panorama regarding the representation of Universals at different linguistic levels. This overview is intended to reflect some of the contributions at different levels to determine whether they are equally reflected. We also want to show precisely what has interested typologists at different levels.

It should be recalled that Chomsky (1965:28) delimited the occurrence of the Universals to “either the syntactic, semantic, or phonological component or to interrelations among the three components”. This selection of levels is, we believe, a full declaration of intentions as to what Chomsky is interested in and what he is not interested in (what he does not make explicit). It is especially revealing that he emphasizes that Universals can be located at an interface between levels. Indeed, we see much potential in such a possibility in Greenberg-type Universals studied in this thesis. Rules that relate, for example, the relationship between morphology and word order (syntax) can be drawn (D. Brown, 2010).

Greenberg (1963:260), on the other hand, delimits the field of action of Universals in: “phonological, grammatical, semantic and symbolic”. In this example, we see the lack of reference to an interrelation of the levels. We also consider the reference to the “symbolic” level problematic since there is no detail, no exemplification, and no reference appearing again throughout his proposals.

However, we want to make a much more general and positive mention of Greenberg’s contribution concerning the (implicit) integration of the different linguistic levels. We should recall that the publication Greenberg (1963) is a book resulting from the conference held at Dobbs Ferry in 1961. In this conference, linguistic Universals were investigated, regardless of the language level concerned. This resulted in a balanced final book with proposals showing a polyhedral view of the phenomenon.

In this book, we find the contributions of different authors and experts at different linguistic levels. In it, they try to respond to the initial concern of Greenberg and his colleagues at the time of convening the conference, which arose from his talks with the psychologist Casagrande between 1958 and 1959 (Greenberg, 1963:iii): common patterns in the different languages of the world. Thus, in addition to Hockett’s initial chapter, where the foundations of human language are laid, we find studies of Universals in phonology, syntax, morphology (inflectional), morphology (lexical), semantics, or morphosyntax.

8.1 Syntactic Universals

In Linguistic Typology, although it has been shown that any level is susceptible to exhibiting Universals, syntax occupies a privileged place, as is often the case in many other linguistic studies (Cinque, 2009). In particular, the Universals based on word order (of syntactic elements) have transcended the most from the cited publication and occupied the most pages (Comrie & Keenan, 1979; Whaley, 2010). Although this thesis will show different examples of syntax-based Universals in detail, we recommend reading Newmeyer (2008) for a more complete and general overview in which deep Universals are considered in more detail.

Greenberg’s proposal was a revolution in Linguistic Typology, precisely with a shift from the morphological to the syntactic. After Greenberg, most typologists have taken it as a point of reference to be able to refine their claims further. This means that much of the effort made in recent decades has been confined to syntax.

Current research on Universals continues to prefer, for the most part, the analysis of syntactic elements. Computer-based typological studies continue to aim to characterize the basic word order of languages, in this case, through new metrics (such as dependency word length).

Evidence of this fact can be obtained by looking at the type of information annotated in corpora, such as Universal Dependencies. In such a case, syntactic information appears, complete with the

part of speech and morphosyntactic features.

However, despite being the most studied linguistic level, syntactic Universals still have a long way to go.

8.2 Morphological Universals

The morphological level, in Universals, has traditionally been in the background, behind the syntactic level. Within morphology, moreover, the most studied aspects are those most closely related to syntax: the features reflected in the different words. Other morphological aspects, such as word creation processes, have been less studied. We recommend expanding these details in the survey of the morphological Universals of Spencer (2010). For an approach more closely linked to Universal Grammar and such deep Universals, we recommend the examples contained in Bobaljik (2012).

Some of Greenberg (1963)'s Universals deal with features such as morphemes and how gender, number, person, tense, etc., are expressed in the words of the world's different languages. These are, therefore, the morphological elements that will be studied the most over the years.

It may be interesting to recall, again, the information that is annotated in large databases such as Universal Dependencies. The features are represented. On the other hand, it is impossible to obtain morphological information that would allow segmenting words (the minimum unit in UD). There are no tags to reflect any morphological process related to words. This ultimately limits the possibilities for investigating morphological Universals of other types computationally. It has taken the last few years for initiatives to emerge that offer relevant morphological information in a resource equivalent to UD. Currently, we have resources such as UniMorph (Kirov et al., 2018). They only have word segmentation in 16 languages. However, we believe that a significant and varied number of languages will be available in the future, allowing for more typologically adequate knowledge.

Furthermore, although morphology is an exciting level from the outset for Linguistic Typology, in many cases, Universals are not studied in a strict sense. It is expected to observe representations of the behavior of the world's languages around some morphological aspects but not under the Universal framework. A Linguistic Atlas such as the World Atlas of Language Structures (Dryer & Haspelmath, 2013) is a clear example of this.

In other words, we can summarize the presence of morphology in Universals with the words of Ponti et al. (2019:5):

Cross-lingual variation applies to all levels of linguistic analysis. The seminal works on Linguistic Typology were concerned with morphosyntax, mainly morphological systems (Sapir 2014 [1921], p. 128) and word order (Greenberg, 1966). This level of analysis deals with the form of meaningful elements (morphemes and words) and their combination. There is also work on phonology, i.e. comparison of elements that are distinctive but meaningless in isolation (phonemes). The systematic comparison of features at these levels is called structural Typology. As an example, consider the alignment of the nominal case system (Dixon, 1994): some languages like Nenets (Uralic) use the same case for subjects of both transitive and intransitive verbs, and a different one for objects (nominative-accusative alignment). Other languages like Lezgian (Northeast Caucasian) group together intransitive subjects and objects, and treat transitive subjects differently (ergative-absolutive alignment).

8.3 Semantic Universals

In semantics, specific difficulties that complicate the generation of Universals are observed in the study. The difficulty of obtaining precise semantic data in multiple languages in a reliable way means that most of the Universals formulated at this level correspond to a methodology closer to Chomsky, based on the in-depth study of a much more limited number of languages (Delfitto, 2009). Evans (2010) explains how the lack of conceptual equivalents across languages and a common framework, such as Wierzbicka (1996)’s semantic primes or semantic prototypes, complicates such studies.

Moreover, separating the semantics from syntactic Universals success is not easy. Usually, there is a use of semantic criteria to be able to subsequently establish comparisons at the syntactic level, something that will be seen to persist today, even in Universal Dependencies (Stabler, 2009).

For further knowledge about Universals in semantics, we recommend reading Von Fintel and Matthewson (2008). This reference provides a thorough survey of the state of the art. We can use this data to exemplify and justify the limitations discussed above. Von Fintel and Matthewson (2008) devote the first part to discussing lexicon-based semantic Universals. The most representative example of this may be Berlin and Kay (1969)’s implicational hierarchy of colors. However, they acknowledge that “we found no useful list of meanings which are Universally lexicalized” (Von Fintel & Matthewson, 2008:190). Moreover, it is essential to note that they devote a large part of the section on lexical Universals to discussing the Sapir-Whorf hypothesis or theory of linguistic relativism. They recognize that, if true, it would go against the existence of semantic Universals.

The second large block is devoted to Universals related to the compositionality of languages. In this case, we observe aspects more closely linked to morphology than semantics. The same happens with what is shown about the Universals related to the morphemes.

The third and last big block corresponds to pragmatics. Von Fintel and Matthewson (2008) understand this level as subordinate to semantics. Even if pragmatics were considered in semantics, no clear contributions are understood as typological implicational Universals. Examples given are, for example: “all languages have presuppositions” or “the cooperative principle” (Von Fintel & Matthewson, 2008:178-190). These pragmatic manifestations are not comparable to Greenberg’s implicational Universals, as they are much more general principles about language.

Ponti et al. (2019:6) summarize the study of semantic Universals as:

On the other hand, semantic Typology studies the semantic and pragmatic levels. This area was pioneered by anthropologists with works on kinship (d’Andrade 1995) and colors (Berlin and Kay 1969). The main focus of semantic Typology is how languages categorize concepts (Evans 2011) in the lexicon, in particular with respect to the 1) number(granularity), 2) division (boundary location), and 3) membership criteria (grouping and dissection). For example, consider the event of opening. 1) It is lexicalized as a single verb in English, but it can be split into a series of subevents in a serial verb construction in Kalam (Highland Papuan) (Pawley 1993). 2) Moreover, it lacks a perfect equivalent in other languages, like Korean, where similar verbs overlap in meaning only in part (Bowerman and Choi 2001). 3) Finally, the English verb encodes the resulting state of the event, whereas an equivalent verb in another language like Spanish could rather express the manner of the event (Talmy 1991). Although variation in the categories is pervasive due to their partly arbitrary construal, languages share cognitive prototypes that constrain it (Majid et al. 2007).

8.4 Pragmatic Universals

If there is one level that we see invisibilized in the study of Universals, it is precisely that of pragmatics. This point of view has also been put forward by other authors such as Habermas (1979), Wierzbicka (1986, 1992, 1994), Trosborg (2010) or Márquez Reiter and Luke (2010).

Besides being scarce, the studies that have been carried out at the pragmatic level have suffered from a rather Anglocentric perspective (P. Brown & Levinson, 1988; Leech, 1989; Wierzbicka, 2003). Other reasons that hinder their study is the unresolved relationship with semantics (Hopper & Thompson, 1993; Meadwell, 1994; Pérez Hernández, 1999; Rama Martínez, 1996).

The two pragmatic aspects that have been tried to be reviewed more from the point of view of Universals have been speech acts (Blum-Kulka, 1981; Blum-Kulka & Olshtain, 1984; Levinson, 2011; Roque, Tsuetaki, Sarathy, & Scheutz, 2020) and politeness (Escandell-Vidal, 1996; Kasper, 1990). In both cases, the results are not as satisfactory as the authors expected because of the

difficulty of working with indirect speech acts in the multiple languages of the world and because they have a high culture-specific content, respectively.

The study of pragmatic Universals is closely related to collecting their data through questionnaires, such as the Discourse Completion Tasks or Tests (DCT). Obtaining these data in multiple world languages with a representative sample requires economic and human resources, making their formulation difficult.

In short, in Linguistic Typology and, specifically, in typological Universals, there is no difference in the treatment given to pragmatics compared to other fields. In other words, we observe a lower representation of pragmatics and its study.

8.5 Phonological Universals

Regarding phonology, we observe less production of Universals over time. In the foundational years of Universals in Greenberg (1963), we find multiple proposals for Universals, and, little by little, we see fewer innovations. Maddieson (2010:1) separates these Universals into three distinct types: “prosodic patterns, segmental patterns, and the structures of longer elements such as syllables and words”. We recommend further reading in Hyman (2008) for a general review of the phenomenon.

Some formulations of Universals are observed in phonology. These are mainly unrestricted, such as: “all languages have stops” or “all languages have at least two degrees of vowel height”. The main interest lies in detecting the elements of the phonological system shared by all languages. Syllabic structures have also been worked on. In the case of Typology itself, there is less formulation of implicational Universals. Where such studies are carried out, they tend to have a form that does not necessarily ascribe to what we understand as a Universal but simply focuses on describing the phonological behavior of the world’s languages of a given type. On the other hand, there is a much more focused avenue of research into diachrony. In this case, it is observed whether invariances in the phonological evolution of languages can be extracted.

9

Universals in The Universals Archive

Contents

9.1	Classification According to the Terminology Distinction	86
9.2	Classification According to the Methodology	88
9.3	Classification According to the Linguistic Level	89

This chapter is intended to be a more quantitative approach to producing Universals in language. Since it is difficult to know the total number of Universals produced, we intend to obtain the figure based on The Universals Archive (Plank & Filimonova, 2000). This is a repository whose main aim is to collect all the Universals formulated throughout history. Therefore, the database gathers the bibliographic information by classifying the Universals and bibliographic information about them.

In order to make the information provided more meaningful and to enable us to detect where most of the Universals are located, we propose a breakdown of this chapter in coherence with the chapters seen previously. This implies, therefore, that we are going to divide the quantities of Universals we offer about:

1. The kind of Universal it is from a terminological point of view. We show the quantities concerning the opposition between deep and surface Universals, implicational and non-implicational Universals, and absolute and statistical Universals.

2. The type of Universal from the methodological point of view. We show the quantities referring to the opposition between Universals from the armchair, the laboratory, and the field.
3. The type of Universal that is from the point of view of the linguistic level. We show the quantities concerning the opposition between syntactic, morphological, semantic, pragmatic, and phonetic-phonological Universals.

9.1 Classification According to the Terminology Distinction

The different productions clearly exhibit an imbalance in relation to the different labels we have seen to qualify the Universal concept.

There is no possibility of contrast between deep and surface Universals in the tool. The tool is made typologically (the one that concerns us in this thesis). Therefore, the total number of Universals housed is understood to be equivalent to the surface Universals. We obtain the total sum of 2177 (surface) Universals. In their general comments, the authors of The Universals Archive do not dare to give the exact amount and mention “over 2000 Universals”. The exact number is obtained by searching the archive without restriction or filter.

The other distinction we have made is related to the presence of exceptions or not. We found the distinction between statistical Universals and the rest of the labels in the classification offered by this repository. There are two statistical Universals in The Universals Archive, meaning that in 2175 Universals, no counterexamples have been found. This is not true. Most of the Universals proposed by Greenberg are statistical, for example.

There is a need for more refined and systematic labeling. Thus, the Universals have not been labeled based on this distinction and have only remembered on two occasions to do so. However, if we forget the “statistical” tag in the “status” section and manually type “statistical” in “quality”, we obtain 827 Universals. If we do the same manual process with “absolute”, we get 1237 Universals. There is also one Universal if we search for “statistical?” and 137 if we search for “absolute?”. We do not know if there are more typing errors or inconsistencies in the labeling because there is no list of the terms used or any guide to guide the user in the manual search. This adds up to a total of 2202 Universals. It should be noted that this result is slightly higher than the total number of languages we have previously presented. This implies that there are several duplicities in the characteristics of some Universals (25) that cause these dozens of extra languages.

However, we cannot guarantee the accuracy of these results. We would face a false case of Universal absolute in any entry in which they have forgotten to add the corresponding exceptions. On the other hand, we can be more confident of the Universals that have been included and contain exceptions since we can be sure that they have not forgotten to include such information.

Another example of the lack of systematicity of the resource can be seen in the existence of overlapping labels. In addition to the distinction between statistical and absolute Universals, they also want to differentiate between diachronic and synchronic Universals, depending on the influence of the time factor on the Universal. We find three labels that present us with many doubts about their differentiation: “unclear whether achronic or diachronic”, “unclear whether diachronic and/or achronic”, and “unclear whether diachronic or achronic”. More order in labeling Universals and a more consistent annotation proposal would make this data available.

The last distinction made is between implicational and unrestricted Universals. In this case, again, we find different labels that may overlap or refer to concepts under the same word and that we must process. We decided to collect under implicational Universals the labels: “implication”, “implication”, “implicational hierarchy”, “mutual implication”, “mutual implication?”, “mutual limitation”, “nested implication”. We decided to collect the labels “no genuine implication”, “non-implication”, and “unconditional” under unrestricted Universals.

These labels either correspond to errors or doubts in their labeling or are more concrete types of implication (e.g., mutual or biconditional). In the case of unrestricted, we include those that present a non-conditional form or a conditional form but which, in reality, is not a condition from the formal point of view (“no genuine implication”).

The implicational Universals would be 1872, and the non-implicational Universals would be 523. The sum of these two groups is 2395 Universals. Therefore, there are 220 more Universals if we perform the search regarding restricted-unrestricted opposition than if we do not establish any filter. We should not stay with the exact quantities, and it is more sensible to highlight the distributions in a general way.

In other words, we can speak of more than 2000 Universals that have been ascertained and stored. Of these, about 60% are absolute, and 40% are statistical. We must remember that in this case, the label absolute is applied if no counterexamples are added to the entry of that Universal; it does not mean there are no exceptions. In terms of structure, more than 75% of the Universals are implicational, and approximately 25% have no preconditions. The sum of these data confirms our suspicion that the high number of absolute Universals is due to an absence of adding

or documenting exceptions and not to a high presence of Universals that apply to the totality of languages in a Chomskyan sense since very few Universals are not implicational.

As a summary, we present figures 9.1 and 9.2, which show the data presented on the statistical/absolute and implicational/non-implicational distinctions, respectively.

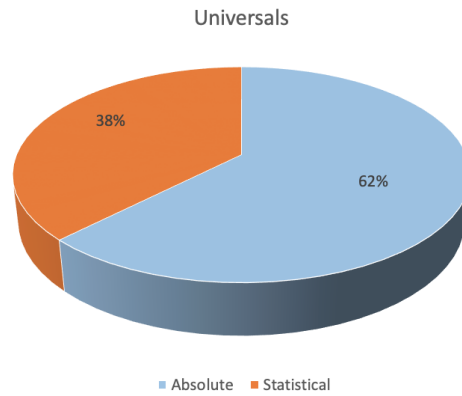


Figure 9.1: Statistical and Absolute Universals Distribution on The Universals Archive

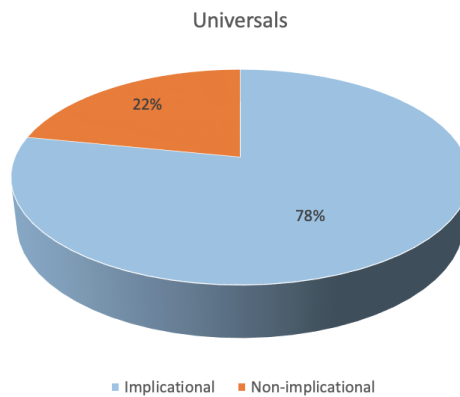


Figure 9.2: Implicational and Non-implicational Universals Distribution on The Universals Archive

9.2 Classification According to the Methodology

Another possible classification of the different productions of typological Universals can be carried out by separating them by the source used to formulate them, as we have seen: armchair, laboratory, and field. In the resource The Universals Archive, there is no explicit reference to this type of distinction. Therefore, a data-based approach to this phenomenon is not possible.

Although the author from whom the data for the formulated Universal has been extracted is cited (most of the time, it coincides with the author of the Universal), it is not possible to carry out a study by manually going to the more than 2000 sources and checking the type of material on which it is based. The problem is that the author's name is referenced, and a hyperlink is provided, which, when clicked, leads to the complete reference for consultation if necessary, but all the hyperlinks leading to the source are broken and do not allow the provision of this data.

In a general approximation, we see that the vast majority of these more than 2000 Universals have been carried out using the armchair method, i.e., by comparing different grammars. It is also possible to find some Universals formulated using a questionnaire, but the number is much smaller.

There are no Universals proposed from corpora. Proposals focus on typological revisions and not on *ex novo* creation. It should also be noted that this type of contribution is still very scarce and very recent, which makes it difficult to include them in this archive.

9.3 Classification According to the Linguistic Level

The Universals Archive also offers the possibility to filter by linguistic level. This can be done by using the drop-down in the search engine under the word “domain”. There are 58 different categories. This high specificity makes it difficult to classify Universals according to their affectation. We often find categories that reflect the mixing of two different linguistic levels. The Universals have been entered without guidance by different people who have decided to use different labels to refer to the same aspect.

We decided to reduce the different levels and stay with the first part of the specification, which results in discourse, inflection, lexicon, morphology, phonology, prosody, semantics, syntax, and word formation. We find 3 Universals on discourse, 777 on inflection, 99 on lexicon, 172 on morphology, 523 on phonology, 54 on prosody, 7 on semantics, 495 on syntax, and 46 on word formation.

In this classification, some interesting data can be observed. The three elements that stand out the most are inflection, phonology, and syntax. Both inflection and syntax are the two most exploited parts in Greenberg's Universals. Therefore, one can see the influence on further development, building on this tradition. It is also necessary to highlight the low number of Universals in semantics found. In this case, we believe that it is due to the limitation of the starting sources. These Universals are fundamentally based on studies through grammars, and in these, semantic

information is traditionally scarce. Likewise, we do not see any entry that refers to pragmatics in a clear way, which goes in the same direction as what we pointed out with semantics, although perhaps in an even more extreme way.

As we can see, the linguistic tradition has a quite evident weight in the Universals and is clearly expressed in these numbers. This tradition conditions the most studied levels because it is the one that nourishes the scholars of the Universals of the data through grammars. Let us hope that, little by little, with the arrival of new means for its study and new tools, this gap can be reduced. However, we must assume that we observe this same trend in the case of computational Universals. In short, it is much easier to trace invariances by crossing hundreds or thousands of data that are much more discrete and concrete (inflection and syntax). Computers are also no strangers to this linguistic tradition and are much better at performing operations of this style. The data used in the various computational models come from aspects such as inflection or syntax, which are more easily annotated universally.

As a summary, we present figure 9.3, which shows the data on linguistic level distinctions in the terms presented in The Universals Archive.

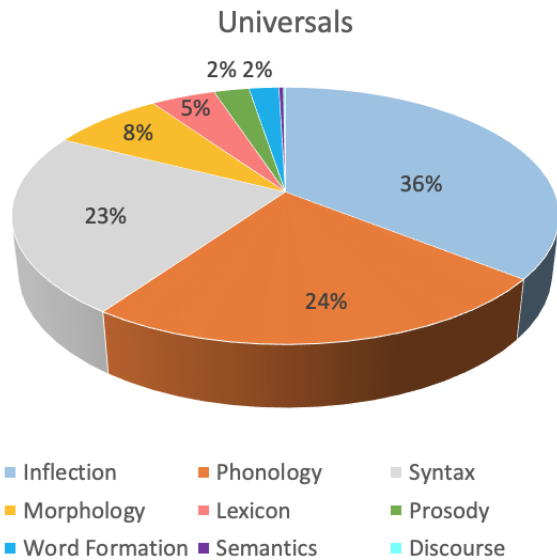


Figure 9.3: Universals Distribution on The Universals Archive According to the Linguistic Level

10

Historical Panorama of the Universals

Contents		
10.1	Universals concerning Greenberg	92
10.2	Computational Universals of Language	111
10.3	Typology and Universals	124

Concerning showing a historical trajectory of Universals in language, we find the division made by Song (2009) to be sensible and coherent. He points out the figure of Joseph Greenberg as the central axis for ordering the periods into which we can divide it. However, Song (2009) proposes a division into four points:

1. The pre-Greenberg morphological Typology
2. Greenberg’s change
3. Greenberg’s follow-up studies
4. The contributions of authors in the 3-4 years immediately before publication (Dryer, Bickel, Comrie, etc.).

We do not advocate making a distinction between points 3 and 4. We see them as a single point. These early 21st-century contributions do not represent any kind of methodological innovation sufficiently different from the publications of the late 20th century to merit such a distinction.

They can also be considered studies after Greenberg and simply present a series of improvements under the same approach.

Thus, as we also understand the American linguist as the major turning point in the subject, we decided to articulate it chronologically:

1. The panorama that existed before Greenberg
2. The turning point of Greenberg (1963)
3. The panorama after Greenberg (1963)

In our case, we also decided to add a fourth point. This corresponds to publications appearing from 2010 onwards. In our case, the time criterion is not decisive for considering a new section, as it is in Song (2009). We consider it necessary to offer a fourth point in which we observe a substantial methodological change concerning the Typology by Greenberg and his successors. The latter, recognizable in Liu (2010), proposes a computational, quantitative, and corpus-based approach to detect the invariance of the world's different languages. Therefore, having completed the historical tracing of Universals in the light of Greenberg, we bring together these main advances in the last division:

4. Computational Universals of Language

We conclude with a historical approach to the different conceptions of the relationship between Universals and Typology.

10.1 Universals concerning Greenberg

First, as we have already mentioned, we will review the study of typological Universals and their relation to Joseph Greenberg's paradigm shift.

10.1.1 Universals before Greenberg (1963)

In the 19th century, when Linguistics could be considered a science, apparent attempts to find Universal patterns in languages began to be observed. However, their development and results differed significantly from what Greenberg would present years later. This does not mean that

the interest in Universals did not exist since ancient times, although it is impossible (and useless), according to Graffi (2010), to establish a precise date for the birth of it all. For example, Hockett (1963) points to Port Royal's grammar as the precursor of taxonomy for Universals. Nevertheless, it was not until the rise of the German linguists of this era that one could find points of comparison with the work on Universals understood by Greenberg. Moreover, in such a publication, the approach to Universals can be considered closer to Chomsky than to Greenberg, given the existing rationalism, according to Bach and Chao (2009).

We will not go into much earlier issues in history either since they are not directly related to Linguistic Typology or Universals. However, they are undoubtedly an exciting postulate in Linguistics. We prefer to refer to the work done by Ramat (2010) and Tavoni (1990), where different, fascinating reflections from the Greeks and Romans are collected.

One of the oldest references is that of von Schlegel (1808), which deals with a first typological separation. Although it does not speak openly about Universals, there is an interest in grouping languages by linguistic type and explaining their characteristics, as Greenberg does in his implicational Universals. Such a contribution can be extended to the whole linguistic school based in Germany in the 19th century, which tries to classify languages typologically according to a morphological criterion (Croft, 2003). Moreover, in these publications, the difference between a typological classification as proposed and a genealogical classification is expressed in an obvious way (Hjelmslev, 1963).

Years later, we find the book by Weil (1844). His study deals with “l'ordre des mots dans les langues anciennes comparées aux langues modernes”¹, as his title already shows. Again, although the concept of Universals in languages is not used, we see very recognizable patterns with the work later to be shown by Greenberg (1963). On the one hand, we see a somewhat similar methodology, with an interest in comparing different languages with variations about different parameters. This is closer to a better representativeness of the world's languages (despite the obvious time limitations). On the other hand, the linguist's topic of interest is the word order, which will be the cornerstone of Joseph Greenberg's proposal.

For Ramat (2010), the main difference between the works of the German school such as von Humboldt or von der Gabelentz and Greenberg lies in explaining why these common facts occur or why such diversity exists. For them, the reasons for this were purely extralinguistic. On the other hand, Greenberg is also open to purely linguistic issues as a conditioning factor. In any case,

¹This could be translated into English as: “Word Order in Ancient Languages Compared to Modern Languages”

we would like to emphasize something more critical than this difference: From the beginning of Typology and the interest in Universals, it has been a discipline concerned not only with a simple taxonomy of languages but also a clear desire for explainability. In summary, we would like to recommend the book by the most significant authority on Linguistic Typology in Italy, Ramat (2010), to expand also this aspect of the history of Typology, with a deep analysis of the first steps prior to the object of interest and study of this thesis. We also positively highlight the parallelism that the author draws between those moments of the 19th century and post-Darwinian botany, as a shared scientific perspective (as well as a similar object of study in terms of behavior) is appreciated.

The following reference to be noted is the publication of Trombetti (1905), an Italian linguist less well-known than the authors appearing in this section. This professor of historical Linguistics at the University of Bologna at the beginning of the 20th century is fundamentally known for defending the theory of linguistic monogenesis. He believed that all languages ultimately derived from a single ancestral language, from which different varieties emerged until we arrived at what we find today. Although there is no mention of all these aspects in Greenberg, and the point of view or purpose is different, the methodology used is very reminiscent of what Greenberg would later use. He tries to analyze grammatical features that different languages of different families may share to trace their common points. He also makes a kind of protocol to review the grammatical rules of the different languages in a balanced way. His research leaves an unmistakable trace of empiricism and the method used to arrive at invariances through a systematic review of the available data.

A few years later, we have a well-known work written by Sapir (1921). Chapter 6 of this work, entitled “Types of Linguistic Structure”, is particularly relevant. It contains what can be considered the essence of Typology, particularly from a morphological perspective. We reference Sapir not because the studies are similar or because we find things that Greenberg later copies but because it is a faithful reflection of how the panorama in Linguistics (or Linguistic Typology) was before Greenberg. This fact allows us to compare and see in what terms the revolution proposed by Greenberg, based on word order and syntax, took shape.

In addition, we also want to rescue what Sapir (1921:60) comments:

The difficulties have been of various kinds. First and foremost, it has been not easy to choose a point of view. On what basis shall we classify? A language shows us so many facets that we generalize from several selected languages. To take, as the sum total of our material, Latin, Arabic, Turkish, Chinese, and perhaps Eskimo or Sioux as an afterthought, is to court disaster. We have no right to assume that a sprinkling of exotic types will do to supplement the few languages nearer home that we are more immediately interested. Thirdly, the strong craving for a simple formula has been the undoing of linguists. There is something irresistible

about a method of classification that starts with two poles, exemplified, say, by Chinese and Latin, clusters what is convenient can about these poles and throws everything else into a transitional type.

As can be seen, in 1921, we already find a clear denunciation of the typical limits of a specific sampling that will appear in Greenberg and be perpetuated up to the present day. The other interesting aspect of this quote is found in the final part. Sapir shows his interest in fixing the limits of the variation of languages in their morphological aspect (or any) and being able to classify the rest of languages within those two limits as a continuum. Sapir’s contribution could be understood as an update and improvement of the criteria and classifications used until then in Linguistic Typology. In any case, the interest is always centered on morphological criteria (Arcodia, 2022a).

Logically, Greenberg (1960) can be understood as an update to this linguist’s work, which is already an update. If Sapir proposes a classification of language with five possible parameters and focuses on morphological questions, Greenberg will take number 4 (“the order of subordinate elements”) and carry out a rigorous and extensive study not previously practiced.

A few years later, we find the fundamental figure of Bloomfield (1933). It is essential to mention that Greenberg explicitly recognizes his work as foundational and is aligned with and inherits these steps taken by the American structuralist in his most famous work. The quote we consider most significant of the relationship we are discussing is found on Bloomfield (1933:20):

The only useful generalizations about language are inductive generalizations. Features we think ought to be Universal may be absent from the very next language that becomes accessible. Some features, such as, for instance, the distinction of verb-like and noun-like words as separate parts of speech, are common to many languages but lacking in others.

These previous words, fundamental for Greenberg’s work, unquestionably commit to the research methodology typical of Linguistic Typology (inductive). They also speak of Universal features, clearly related to the later work of the 1960s. The last key to Bloomfield’s words is the reference to the naturalization of exceptions. As will happen in this field, exceptions to the identified patterns are expected, which does not change the research’s value.

In addition to Bloomfield and Sapir, Greenberg (1960) also explicitly mentions both the nineteenth-century pioneers of Typology (German) and the group of twentieth-century linguists working with a perspective or interest comparable to that to be observed in 1963. For example, he explicitly mentions Jakobson, Hockett, or Mezerath.

Many of these names will be dealt with in the following subsection since their contributions coexist with Greenberg and are included in the same publication. On the other hand, it is necessary

to highlight the group of linguists who dared to formulate laws close to what can be understood as quantitative Linguistics or even mathematical Linguistics. In the case of Menzerath, we highlight the law of the same name. To summarize, an increase in the length of a sentence, for example, leads to shorter clauses. That is, there is an inversely proportional relationship between the size of a linguistic construction and its constituents. In the case of Zipf (1935), a law is proposed in which the most used word (of a language, a corpus, etc.) will be twice as frequent as the second most used, three times as frequent as the third most used, and so on. In these cases, we see laws that, although abstract, have a clear and demonstrable empirical basis and a Universal nature. This will be reviewed in the Ferrer i Cancho (2003) Doctoral Dissertation almost one century later.

In addition to the references cited by Greenberg about quantitative studies to classify languages, we found some others. We do not know whether the lack of reference to these is synonymous with ignorance of their existence. An example of this is the proposal by Krámský (1959), who shows how it is possible to classify Slavic languages thanks to frequencies in the phoneme inventory of vowels and consonants.

Just after the 1930s, in the words of Cooreman and Goyvaerts (1980), American Linguistics exhibited the opposite interest to that of Greenberg. The linguist's interest will focus mainly on emphasizing the diversity existing in all the world's languages instead of trying to achieve common patterns in all of them (or in groups of them). The growing interest in Native American languages and their linguistic descriptions boosted this perspective. All this means that there are few contributions in the years prior to 1963 that can be mentioned (beyond what has already been said in this section) and that, to some extent, the path proposed by Greenberg is even more revolutionary.

That is, the perspective of anthropological relativism prevailing in the society of the time leads to a view that the world's languages can vary in an unpredictable, unknown, and arbitrary way (Croft, 2003). Therefore, the revolution to be seen in Greenberg is a proposed search for invariances within the variation of languages, that is, for trends in this variation or constraints in this variation. This shows that languages do not vary infinitely, and such constraints represent Universals in Language.

The last great figure that we want to highlight in this period relevant to this thesis for multiple reasons is that of Tesnière (1959). We do not know if Greenberg was aware of the cited contribution of the French linguist, better known for the valences. This is a case of fame and significance of this work that has occurred more than half a century later. The study focused on interesting aspects of

Linguistic Typology and was published only in French and posthumously. Moreover, according to Osborne and Kahane, the translators and editors of the 2015 edition (Tesnière, 2015), the temporal coincidence with Chomsky and his premature death also add to reasons why his contribution and vision of syntax have been silenced to this day, even though some of the interests that appear in Greenberg are already seen in Tesnière. We find, years later, translations in German (Tesnière, 1980), Russian (Tesnière, 1988), Spanish (Tesnière, 1994), Japanese (Tesnière, 2007) and Italian (Tesnière, 2008). However, it is not until less than ten years ago that we have an English translation (Tesnière, 2015). Although the idea of dependency can be found in earlier contributions by Stephen Clark half a century before him, for example, he is the first to develop a complete theory of the phenomenon.

In his study, we explore the idea of classifying languages into two fundamental types: centrifugal languages and centripetal languages. Although it seems simple, such a distinction will be fundamental for some of the recent studies that are being carried out opposing the types of head-initial (equivalent to centrifugal) and head-final (equivalent to centripetal) languages. Although it is a rather general and generic classification, we value it as an alternative typological classification proposal, which, moreover, is very fruitful and was produced before Greenberg.

We did not find the information we will comment on until 2015, when Osborne and Kahane published a complete and translated edition into English (Tesnière, 2015). We find a vast and exciting documentary research work in the foreword to this version. These two linguists have accessed boxes with the author's notes (from 1936) in custody by his daughter and never analyzed. In them, they have found, among many other relevant things, a typological study on word order, classifying around 190 languages with parameters such as “noun-genitive”, “noun-possessive”, and “verb-subject”. The embodiment of this research can be seen in chapter 14 of the book Tesnière (1959) and shows clearly an almost perfect equivalence, not only with the work that Greenberg will carry out years later but with many linguists who have come later.

10.1.2 Universals in Greenberg (1963)

We have previously highlighted authors who pointed out how it is impossible to determine a clear beginning of Linguistic Typology. However, we can put a specific date to this discipline's most crucial turning point. We refer to the change of conception from the morphological to the syntactic criteria, the use of precise empirical methods for formulating such theories, and the birth of rules known as Universals in Language. This date is, symbolically, 1963.

The publication containing this revolution is a conference proceedings; therefore, reference can also be made to the dates April 13-15, 1961. In this period, Greenberg organized a conference in Dobbs Ferry (New York). Such a conference managed to interest multiple reputed researchers, such as Jakobson or Hockett, who also participated in the publication. We also know that the theme for the conference emerged in the academic year 1958-1959. It was commissioned by Joseph Casagrande, a member of the Linguistics and Psychology Committee of the Social Science Research Council, to the three committee members who were Stanford researchers that year: Joseph Greenberg, James Jenkins, and Charles Osgood.

The initial idea was that these researchers would draft a memorandum (Greenberg et al., 1963) that would include all Universals that could be detected for the world's languages. This list was then to be discussed at the conference. The researchers at the conference were to complete the study from different perspectives (given their backgrounds). Joseph Greenberg assumed the most significant responsibility in such a conference. He made a more transcendent and clear scientific proposal, which has caused him to stand out compared to the rest of the participants in this publication. His commitment to this subject, his work in charge of the whole process, and his outstanding personal contribution to the book make him the visible head of this revolution in Typology and Linguistics.

Universals as Features of Language. Thus, at this particular moment that we isolate in this subsection (the 1963 publication), we are eager to offer a representation of Universals at any linguistic level. As we have already commented, by introducing the subject, we find a distinct conception of Universals from Hockett (1963), which shows the limits of human language without mentioning the relation to other aspects that ascribe its type. In Chapter 1, we find a transcendental and fundamental contribution in Linguistics, which goes beyond the study of Universals. In any case, it is necessary to comment that we are dealing with a very general proposal that does not facilitate the appearance of exceptions and allows us to establish an initial common framework.

Universals of Linguistic Change. One of Greenberg's interests, which we know from his foreword, is to bring order from a diachronic perspective. The second chapter, by Hoenigswald (1963), accounts for this. The perspective headed by Greenberg assumes the wide variety in the world's languages and understands that this diversity is mutable in many respects. Linguistic change, until then, had been described, in most cases, simply in isolation (or by families). From this moment on, we find an interest in systematizing this change in all the

world languages, starting from the type. In other words, it is understood that the behavior of languages is always the same, even though they may change or be different.

Phonological Universals. In the third chapter, we find the contribution of Ferguson (1963), who focuses on nasals. The phonological Universals he proposes are pretty general despite the specificity of the subject, a fact that does not allow a grouping through different types as can be done in Greenberg. Something exciting and striking about the proposal is the methodology used. It is possible to appreciate the concerns that, from that year on, will be essential in this discipline, gathered under a typical sample study. Saporta (1963), in chapter 4, also deals with aspects related to phonemes and the world's languages.

Greenberg Universals of Word Order. Subsequently, we find the star research of the publication, in chapter 5, in which Greenberg (1963) formulates in a clear way grammatical Universals (and highlights in the title, especially those based on word order). This proposal is also based on the methodology followed in all subsequent years to carry out studies of this type (although improving or extending some aspects). Greenberg's selection is of 30 languages, known to him or through grammars. He assumes that it is not a perfect sample. There is a particular interest in promoting balance and representativeness in the sample that allows us to approximate the behavior of the world's languages. Although in this publication, there is also an interest in following a rigorous research methodology in some colleagues (not all), the method and description of this observed in Greenberg far surpasses any other attempt because of its methodical approach.

Therefore, Greenberg creates different linguistic types around a given element and tries to see how many languages of each type exhibit a pattern. In the case of Universals in which correlations are established in word order, it will always be seen how there is an implication. That is to say, there is no single Universal order, but rather, once a language is ascribed to one of the possible orders, it manifests particular behaviors regardless of its family or area. This type of Universals within the list presented by Greenberg are the ones that have transcended the most within the field and have been considered essential for the revolution in Linguistic Typology that has taken place.

Most Universals are based on a review of linguistic types about the basic word order of nominal subjects in co-occurrence with verbs and nominal objects in unmarked sentences. The end product (mainly SOV, VSO, and SVO) has been fixed and naturalized as basic information for any linguist about any language. This concern for the order of constituents as a replacement for other different criteria prevailing in Linguistics was the real revolution.

There is no such revolution in Universals that refer to questions of another type, such as the presence of a morphological element if another existed previously. The more general data have not aroused as much interest as the previous Universals. In this case, we can emphasize that the methodology used is the same, in clear contrast to proposals that may arrive at similar-looking results but without such clear traceability.

Universals and Indo-European. Later, in chapter 6, we find a study on Universals in Indo-European languages (Cowgill, 1963). In this case, we must point out that it is a misleading title since it is not possible to equate a Language Universal to what genetic characteristics are within the same group. The conclusions do not imply scientific knowledge relevant to the field of Language Universals.

Semantic Universals. The following two chapters (7 and 8) are devoted to the semantic part (Ullmann, 1963; Weinreich, 1963). The results that can be observed are neither transcendent nor significant. In this case, we believe that the fault does not lie so much in the lack of methodology on the part of the authors but rather in the shortcomings of the sources. The semantic level does not enjoy much importance in describing the grammars of the different languages. Because of this limitation, it is challenging to work with it. At present, as we shall see, this limitation is very clear. However, we continue to see an underrepresentation of semantic data in a format comparable to the morphosyntactic one for studying typological Universals.

Universals and Antropology. Casagrande (1963), in chapter 10, deals with the relationship between Universals and anthropology. It focuses primarily on seeing whether culture has an influence on all these facts and whether there are Universal traits in it. Although it is a necessary chapter because of the traditionally neglected nature of this field (although the relationship between anthropology and languages was very fashionable in the United States at the time), no revolutionary conclusions are observed. We understand that this is due, fundamentally, to the fact that the methodology used by Casagrande is the traditional one prior to Greenberg.

Universals and Psychology. Osgood (1963) proposes in chapter 11 how the paths of Linguistics (as it has been proposed in this conference) and psychology are compared. Here again, it is a reflection in which no new Universals are presented, but which allows us to understand aspects that we take for granted today and were also revolutionary. More than a few linguists today speak of the explanatory power of Universals and the functionalist concern behind any

claim in Typology. This means that the linguist's interest is not only to detect Universals but also that these are the previous steps to understanding some kind of law that forces the existence of these Universals.

The other aspect that Osgood explains perfectly and is fundamental is about exceptions. Again, we find some linguists who explain the possible cause of the exceptions present in some Universals. The explanation would be based on the understanding that the exceptions reflect the interactions between the whole set of functional laws. That is to say, different laws may come into play simultaneously in a given element. In the case of contradiction, one of them conditions the chosen form, a fact that causes another of the laws not to be fulfilled and generates such an exception. Much commented on in many recent papers, this fact is already found in Osgood (1963), although it appears cited in a few bibliographies.

Reflection on Universals. Jakobson (1963) is the author of chapter 9. This chapter does not show any original output since it focuses on carrying out a necessary reflection on the contributions made that year about Universals in language, with particular attention to Greenberg's proposal. The paper's central idea is based on the implications of a revolution such as the one proposed in this conference. In general, we have also presented all these facts in this thesis. The most relevant aspect for Jakobson is the study methodology behind Greenberg's Universals. He believes that, although the output may be more or less questionable, in the future, thanks to improvements in data availability or techniques, these Universals may or may not be validated, and new ones may be born.

Finally, by way of closure, we find the memorandum (as an Appendix) (Greenberg et al., 1963). Rather than offering the Universals, something that does appear in Greenberg's chapter (we understand that this is because he is the one who carried out all this research), it lays the foundations of the new version of the discipline being born. The three authors responsible for the memorandum devote themselves to presenting a series of concepts and labels (and defining them) that are fundamental to this study (e.g., "Universal statistical"). Thus, what is done is to provide all the necessary tools so that complete research can be carried out with the most significant possible autonomy.

In short, what Greenberg (1963) contributes to science goes far beyond the 45 statements that he develops, which we will later see in detail and analyze. It is necessary to emphasize his work as a pioneer of the work of his colleagues, which is also based on Greenberg (and the memorandum). In addition, the American linguist emphasizes that the revolution of his research is to be found in "the establishment of Universals based on the empirical linguistic

evidence” (Greenberg, 1963:59-60). All the methodological work that allows a different approach to facts of interest for Linguistics and that allows, in the following years, a refinement of the proposed models is important. As a matter of course, the exceptions that he accepts are also collected rigorously and systematically at the end of his paper, an uncommon transparency for that time. Another of the positive aspects of such a methodology is that it allows research at any linguistic level, even though Greenberg focuses on morphology and syntax. Another of the revolutions in the field, as has been mentioned, is the change in the weight of both levels, giving the syntactic one a more significant entity, contrary to what happened before Greenberg in the field.

The entire scientific community has recognized all these issues. Thus, we find authors such as Comrie (1989), Christiansen et al. (2009), Song (2009), Ferguson (1978a), Pericliev (2018) or Arcodia (2022b) who explicitly recognize that without the figure of Greenberg, today, we would be talking about something completely different and their work could not have been born.

10.1.3 Universals after Greenberg (1963)

After the revolutionary publication mentioned above, a large number of studies emerged. In general, to extend this section and to better understand the revolution that Greenberg brought about with the previous linguists, we recommend Haspelmath (2009) and to understand better the revolution that Greenberg brought about through the linguists that came after him we recommend Scalise, Magni, and Bisetto (2009) and Song (2009). However, we propose a review structured by the main actions that have been carried out:

1. Discarding the subject as a significant element.
2. Proposing different and new Universals.
3. Proposing a better sampling to study Universals.
4. Offering new applications of Universals.
5. Creating databases.
6. Reviewing Universals.

Discarding the Subject as a Significant Element

In the decade following Greenberg (1963), we observe several proposals that try to take Greenberg's one further through different modifications. Vennemann (1974, 1976) questions the cornerstone of Greenberg's type proposal based on the primary order discussed above. For Vennemann, it is unnecessary to investigate the basic three-element word order since the subject order has no bearing on anything. Therefore, he proposes a much more basic distinction based on the opposition object-verb/verb-object OV/VO (called, by him, operands and operators). That is to say, creating an analogy with what will be seen later as the opposition between head and dependent, he proposes that languages that place the dependent (the object in this case) in a postposed way will place the rest of the dependents in the same way.

This list of elements influenced by the position concerning OV/VO is refined by Lehmann (1978) a few years later. This linguist also believes that it is more convenient to forego taking the subject into account. In addition, he offers different arguments: in many sentences, it does not exist; in many languages, it can be encoded in different ways, etc. This proposal, often understood as a joint one of Lehmann and Vennemann, has subsequently been widely exploited and has proved quite productive. It should be mentioned that although typologists have accepted this new classification proposal well, there is no clear consensus about it. Reputable current linguists such as Newmeyer (2007) still see Greenberg's original proposal as much more fruitful and successful.

We continue to find new proposals for classifying the world's languages. However, they do nothing more than confirm Greenberg's revolution based on word order, breaking with the Typology based on morphology holistically (D. Brown, 2010). Nichols (1986) proposes to understand languages as head-marking or dependent-marking, depending on whether the morphological information falls on one or the other element. For example, we would have both options in English: *Laura's Book* and *the Book of Laura*. She tries to establish correlations with the verbal position and other elements from this typification.

Subsequently, we also find much more global and reflective publications, which try to understand and show the path of the Universals, starting from Greenberg and reviewing the different proposals or modifications throughout the following years (Siewierska, 1988). The various inaccuracies detected and the implications of their subsequent revision, thanks to data from more languages or slightly different approaches, are denounced.

Proposing Different and New Universals

One year later, we find another great revolution based on Greenberg's implicational Universals, thanks to Comrie and Keenan (1979). The British linguist builds on the invention of implicational Universals to take it further and propose an implicational hierarchy. That is, he can propose new implicational Universals, not in isolation but interconnected, which makes a series of implications dependent on the previous ones. In his case, he proposes the so-called "Accessibility noun phrase hierarchy". In this, Comrie and Keenan (1979) can establish a hierarchy of syntactic elements that can be relativized. For example, if an indirect object can be relativized, a direct object and a subject can be relativized. Such complex Universals allow a compelling reading from a variationist point of view. It is possible to predict the next step of a language (gaining a step or losing a step, but in an orderly fashion). It also applies to different language varieties, which may allow some more or fewer steps than other varieties but always follow the order marked in the hierarchy. Sigurd (1989), a decade later, tries to apply this hierarchy to different languages and study the more complex cases behind Comrie's study. Other hierarchies have been born, for example, in Luo (1994). In any case, their relationship with Greenberg is not so direct, and they owe him the methodology used for the implicational Universals, which we have already presented in the previous sentences applied to the hierarchy shown.

Proposing a Better Sampling to Study Universals

We can highlight the contribution of Pym (1979). His paper focuses on presenting a new grammar for an unattested language. In this case, Greenberg's influence can be seen in two fundamental respects. First, there are many mentions of terminology or aspects Greenberg worked on. That is, the grammar is made thinking of its usefulness for future studies of Universals understood with Greenberg's innovations. Secondly, the author seeks to offer counterexamples to the previously formulated Universals, in this case, in the Iwaidja language.

We also find room for a publication of Greenberg, Ferguson, and Moravcsik (1978) in which they give an account of these advances or possible improvements on the original 1963 approach. Fundamentally, we can categorize it as a confirmation of his theory and ratification more than a decade later.

A year later, we find a revision to some of Greenberg's Universals by Hawkins (1980). He tries to offer more details of some languages to review these proposals quantitatively. The main difference that can be highlighted concerning Greenberg is that Hawkins gives much more importance to the

position of adpositions as a basis of typological organization. Subsequently, we find the publication of Coopmans (1984) as a critique and revision of Hawkins' proposal. In this case, there is a greater distance from Hawkins's proposal and a closer approach to Greenberg's original one.

Other studies such as Z'graggen (1980) also seek to expand the selection of languages (in this case in Papunesia) to try to qualify or extend Greenberg's Universals to languages that can potentially show deviations from the Universals (because they have not been taken into account in the original set).

The eagerness to extend languages and to draw Universals leads some authors to simply collect in an orderly way such characteristics of the different languages of the world. Hawkins (1983) offers a selection of typological data for more than 300 languages so that they can be considered when drawing correlations between them. Here, the influence that Greenberg has (and will have) can be clearly seen.

Some years later, we also observed proposals that assume Greenberg's results to be generally good and share his methodological proposal but believe that the selection of languages used should be improved to have greater reliability. For this reason, they focus on proposing improvements in the sampling to allow better representativeness. A clear example is Tomlin (1986). In this case, thanks to it, we finally obtain a precise representation of how Greenberg's basic word order is distributed in the languages of the world according to Tomlin: SOV 44.78%; SVO 41.79%; VSO 9.20%; VOS 2.99%; OVS 1.24%; OSV 0%.

We also find much more precise publications, which point out specific aspects of Greenberg that are questioned and shown to be erroneous thanks, precisely, to a correct and extensive sampling. An immensely clarifying example is that of Dryer (1988a). In this case, he focuses on disproving the relationship between OV (or SOV) languages and adjectives preceding the noun they modify and vice versa with VO languages. He shows no such correlation with an analysis of 316 languages.

In other papers by the same author (Dryer, 1987, 1988b), we find pretty large samples (a maximum of 542 depending on whether the linguistic element is sufficiently extensive in the languages). This allows the typological precision to be much higher and the representativeness more reliable.

However, the greatest revolution of this author is found in Dryer (1989). In this case, we find a methodological innovation sufficiently large to be highlighted autonomously: the macro-areas. Dryer improves the sampling of languages to analyze questions of Universals or Typology and orders it differently. The linguist creates a series of macro-areas corresponding to independent

and unconnected regions from a historical perspective: North America, South America, Africa, Eurasia, Papunesia, and Australia. The different languages, comprising genera or families, are placed in their respective macro-areas. In this way, the Universals or their exceptions can be seen more coherently depending on their presence in all the macro-areas.

The next innovation of this author comes a few years later (Dryer, 1992). He summarizes the theories about the Universals governing the behavior of languages and their basic word order in 3 groups. First, he highlights Greenberg's proposal. Secondly, the Head-Dependent Theory (HDT) stands out, which is the proposal previously seen by Lehmann and Vennemann. Thirdly, we would find Dryer's new proposal called Branching Direction Theory (BDT), which he will consider even in his latest publications of a few years ago (Dryer, 2013).

This BDT is very similar to the HDT proposal. The main difference lies in the elements that are placed in each group. If the two linguists of the 1970s equated the behavior of relative clauses to adjectives as having similar syntactic behavior, Dryer (1992) will offer a different reading. Only phrasal elements can be grouped with other phrasal elements; the same is true for non-phrasal elements. On the other hand, Dryer goes further than Lehmann and Vennemann and confirms that it is not simply that Greenberg's proposal about adjective order is wrong, but that he demonstrates with quite evident data that it is a non-correlation pair, that is, that the behavior of this cannot be predicted. In the case of correlations of different elements, he wants to verify them and understand why they occur, which is essential for this functionalist view of the functioning of language. This view will be again in a similar paper in Dryer (2009).

The improvements in sampling previously shown in Dryer will be the most worked theme in Typology in the 1990s. Rijkhoff, Bakker, Hengeveld, and Kahrel (1993), for example, will also try to offer a better balance in sampling. The differential aspect is that they are interested in balancing any genetic influence on the final language selection to reduce this bias.

Another aspect launched by Greenberg, which is fundamental for his theory and later contributions, is that of dominant word order. It is also the subject of analysis and discussion at this time, especially in Dryer (1995). This interest will also be observed in later publications, such as Dryer (1997). The linguist questions the criteria used by Greenberg to determine it. In his opinion, more weight should be given to issues such as markedness or other elements relevant in languages not envisaged by Greenberg. An aspect that Dryer does not point out but that we add in this sense is that if Greenberg relied on order frequencies to determine which was dominant, he could not count on concrete quantitative data that would allow such a statement. Given the nature of his

sources, in which he relied on the label present in a grammar or his knowledge of it, a threshold for this cannot be objectified.

Bickel (2007) points out two fundamental changes in this post-Greenberg landscape: “Universal areality” and “advances in methodology”. The first of these refers to the Dryer-driven revolution to consider macro-area aspects when validating Universals. The second refers to improvements in sampling and using quantitative techniques to validate claims more accurately. In any case, as can be seen, beyond these two specific aspects, the Greenberg revolution is still going strong and is a fundamental reference in the field. Moreover, it is essential not to overlook that, as Song (2009, 2010) points out, all post-Greenberg proposals from the mid-1980s to the present naturalize and assume statistical Universals as important.

From then on, a way is opened to new methodologies or representations of Universals that allow a better understanding of those already formulated or the proposal of new ones. For example, we find the Multidimensional Scaling proposal of Croft and Poole (2008b, 2008a). This is an interdisciplinary work in which innovative techniques used in mathematics and statistics are applied to provide a perceptual mapping (a visual diagram) that offers a quick visualization of the behavior of languages. For example, such a tool, often used in marketing, allows for a more comprehensive and accurate data layout.

Creating Databases

In the 21st century, interesting databases and proposals from the statistical and quantitative point of view are beginning to appear that try to improve the verification methods used in Typology (Cysouw, 2003). We must highlight two of the databases: The Universals Archive and The WALs.

The Universals Archive (Plank & Filimonova, 2000), as we have seen, is a project that tries to store and put order to all the Universals proposed in the world. The relationship with Greenberg is clear: without his proposal for formulating these laws known as Universals, there would not be such a proliferation of other Universals that would generate the need to create such a database. The ease of access to the archive (online) and its very intuitive use should be noted. The search engine and the order in which they have been stored allow us to know the Universals at each level and search them more efficiently.

The WALs is the acronym behind the World Atlas of Language Structures, a product linked to the Max Planck. In this case, it was conceived as a paper version (Haspelmath, Dryer, Gil, & Comrie, 2005), but the web version is the one that has triumphed and transcended (Dryer

& Haspelmath, 2013). Although the original idea was to have a linguistic atlas with different structures, different typologists worldwide use this tool for their studies, only to compare these data without taking into account any map. There are many languages, and the data were obtained by consulting a grammar for each language. There was also a Russian version of WALS (Jazyki Mira), which we do not discuss in detail because it was more of a project than a reality, having fallen into disuse shortly after its creation (Belyaev, 2008; Moldovan, Skorvid, & Kibrik, 2004). The main differences lie in the size: 300 in the Russian proposal and more than 2000 in the case of Max Planck’s product (Polyakov, Solovyev, Wichmann, & Belyaev, 2009). Also, the weight of Eurasia, in Jazyki Mira, is higher because it has fewer languages. However, there are quite a few incomplete languages in WALS, with only a little information (Cysouw, 2009). Another criticism that can be made of WALS (or any of these databases) is the reductionism behind the assignment of a label, according to Plank (2009): it appears that both English and Russian are two SVO languages. For the author, it is essential to point out the rigidity of the SVO order in English and the flexibility but dominance of SVO in Russian, a fact that offers crucial information about the behavior of both languages and that tends to be silenced. Something also denounced by O’Horan et al. (2016), exposing that the label for French is that of an adjective behind the noun it modifies, although there are many cases where this is not the case. The controversial limitations of databases such as WALS will be discussed in detail below, although Ponti et al. (2019) is also a recommended reading.

Reviewing Universals

Another aspect we can criticize is that Greenberg Universals are formulated in natural language. This criticism, which is not seen in the bibliography but which we do present, points out the possible ambiguities or inaccuracies that sometimes lie behind the expressions in natural language. There are many examples. For example, we do not know what application the different quantifiers used by Greenberg have and whether there is some kind of logic behind them. What is the difference between a Universal in which the expression “almost always” is displayed and the expression “with overwhelmingly more than chance frequency”? Another possible confusion that the linguist may have is the application of the elements present in the Universal. When Greenberg mentions, for example, the position of the adposition, we do not know what exactly he meant. A reasonable reading may be to understand that he refers to the canonical relation of adpositions, that is, with nouns. A much more extensive interpretation may refer to adpositions with any grammatical part of speech (such as adposition with a pronoun).

In the banal example above, no significant changes in the possible outcome are presumed. However, this fact may have different implications depending on what it applies to. The lack of explicit mention by Greenberg is easily solved by adding a higher degree of precision and using expressions that are not ambiguous. We understand that one possible way to do this is what we propose in this dissertation: converting Greenberg's natural language to formal language. Other authors, as Stabler (2009) exposes, have explored the possibility of converting these Universals to other linguistic schemes, although without much significance.

This same year, we found many other relevant contributions to the discipline, thanks to the revolution proposed by Greenberg. The book promoted by Christiansen et al. (2009) gathers different contributions from multiple perspectives that try to offer an approach between the Universals understood as Greenberg and Chomsky do. Despite the transparent generativist background of this publication, it can be seen how Greenberg's proposals are well received and used as a starting point to improve within postulates closer to Chomsky.

So far, we have only seen, on most occasions, studies that try to improve Greenberg's language sampling or question some general aspect of his methodology to achieve better results. However, we are also beginning to find some studies that try to review in a much more precise and concrete way some of Greenberg's Universals with new data (Cinque, 2013a). In this case, Universal 23 is revised with a new sampling that is much more adequate and disproves the order of the genitive proposed by Greenberg.

Something similar happens in Merlo (2016) about Universal 20. The study collects data from multiple papers (even if they had another objective) such as Dryer (2006), Cinque (2005) or Cinque (2013b) to use them for the revision of this Universal. This aspect of reusing data from other authors looks pretty sensible. In addition, we also find a view of the phenomenon very interesting: Merlo (2016) points out that the interest in defining only the impossible and the impossible in languages has been overcome and is moving towards a more fine-grained approach that allows predicting the frequencies of other distributions. Despite this accurate reading, we miss a framework or model to determine the validity of these quantitative distributions.

In the following years, we also began to find criticisms of these new proposals that were born to solve some inaccuracies in Greenberg or the evolution in Dryer. This is the case of Dunn, Greenhill, Levinson, and Gray (2011), who question the data collected in the different studies based on BDT and HDT. The innovation in this paper is mainly methodological since some families are determined, and this distinction is irrelevant. One aspect that can be criticized in this paper is

that it extracts the data from the WALS, and, therefore, in addition to the existing limits in this database, we find a repetition of the most commonly used sampling. As a result of this criticism, we find, in the same year, the response by Dryer (2011). In this one, an attempt is made to respond with the data of the languages that have been questioned previously. In some cases, Dryer accepts such criticisms in a veiled way. This can be seen, for example, in the quote: “While I did treat the order of noun and demonstrative as a non-correlation pair in Dryer 1992, it is a borderline case. Since my database has grown over 150% since my 1992 paper, it is worth revisiting this pair of elements” (Dryer, 2011:347).

In Kirkegaard (2016), we find the same idea as in Dunn et al. (2011). There is a critique of Typology’s lack of statistical criteria, and they resort again to the labels present in WALS. Similarly, Guzmán Naranjo and Becker (2022) work with data extracted from WALS (hence, grammars) to propose new claims based on Greenberg Universals. For example, one of the main findings is the claim that the order of direct objects and verbs does not influence the preference for affix direction choice. One of the authors, Becker (2021), will do the same by trying to classify typologically the different types of articles in the world’s languages and extracting invariances. However, she will rely on her sources this time and will not reuse WALS.

In recent years, we have found papers that follow the same precepts as those we have been showing so far, as in the case of Dryer (2019). We believe this aspect, which may seem obvious and dispensable, is quite illustrative given the next section that we will present. Although we have observed a boom in studies using a corpus in Typology in the last decade, this does not mean that the approach to Universals or Typology through sampling grammars has fallen into disuse. In the cited paper, we observe a study based on 576 different grammars of languages. It is also important to note that this is convenience sampling, i.e., no corrections are made, and any available data is analyzed. In this case, the relationship between adjectives, numerals, and determiners, which corresponds to a Universal formulated explicitly by Greenberg, is studied. The primary purpose of the study is to try to understand why there are so many exceptions with the order of adjectives, and the conclusion is reached that, perhaps, a new reading is needed on what we consider an adjective or not in the languages of the world, an aspect that we will deal with later on. Another example of this (specifically about the numeral system in this case) through grammars in our days would be Dryer (2019): it starts from Greenberg’s Universal about numerals to be able to go further and offer information about the procedures of addition, subtraction, numeration of languages.

Another of the innovations that we found in that year is, precisely, the publication of Masini and Mattioli (2019). In this case, they apply all the knowledge about Universals (thanks to Greenberg)

to other elements traditionally undervalued or ignored in grammars (such as lists). In this case, moreover, we see a middle way concerning the sources: the data are firmly based on grammars, although the authors open themselves to a later revision using other methodologies to confirm these claims.

Another possible avenue of research generated by Greenberg is the opposite: to review the typological features he proposed as uncommon. Schnell, Haig, and Seifart (2022) try to justify, through a statistical approach, why limits are observed in the variation and appearance of the basic word order OSV since it violates two fundamental laws, the first of them a Universal of Greenberg: subject before object, and object adjacent to the verb.

In 2022, we find a concern for representing Universal-dependent data more visually, quantitatively, and modifiable. It can be done through a corpus approach, as we will see in the following section and this thesis, but also through grammars, as we have mentioned in this section. The clearest example of this is Sansò (2022), who proposes an adaptation of semantic maps to represent the behaviors of languages or with multidimensional scaling.

On the other hand, despite the large amount of work that Greenberg’s typological revolution has given rise to, as Bickel (2010) points out, there are a large number of Universals proposed, but very few of them have been tested in a rigorous or alternative way.

10.2 Computational Universals of Language

In the 21st century, we find gradually a growing interest in uniting the most recent advances in a trendy methodology such as linguistic analysis through corpora and the typological study of Universals (Gries, 2009).

First Steps before Liu (2010)

This interest can already be seen in papers in which perhaps the weight of Typology was more questionable (since only three languages are analyzed) in Xia et al. (2001). The main problem of these first proposals, besides the insufficiency of languages and the non-adoption of typological criteria, lies in the problems of comparison observed in the sources. The chosen corpora have been annotated in a completely different way, making this exercise difficult (O’Keeffe & McCarthy, 2013). We do not point this out as a shortcoming of the paper, as there were not many alternatives, although it should be noted.

These initial moments, in which Corpus Linguistics is becoming fashionable, but the appropriate means are not available, also serve for more reflexive contributions such as Steinvall (2002). This paper discusses the advantages of studying language behavior by adopting corpora. However, it points out the many existing limitations or methodological problems: how to collect the material, obtain precision, etc.

Therefore, in the following years, we will find this interest in improving the work with corpora. For example, Martí, Taulé, Bertran, and Màrquez (2006) show the process of corpus tagging in Spanish and Catalan. Even though the consistency of the labeling is starting to improve, it is still difficult to compare if no guide unifies the labels or if the grammatical approach to syntactic analysis is not shared and fixed (Dryer, 2009).

One of the apparent problems of corpus preparation is the large size of some of them. Therefore, it is logical to see the appearance of papers in this era that also began the search for automatic annotation of corpora for use in Linguistic Typology. In the case of Lappin and Shieber (2007), an effectiveness between 60% and 70% is achieved. The main problem is that it only works in English. In our opinion, the most relevant aspect of this contribution is the effort to improve automatic labeling with precision for typological studies and the benefits in the opposite direction. That is, how useful it would be to configure the linguistic types of the languages to be labeled to facilitate the work of automatic labeling. We are referring to morphosyntactic tagging. The knowledge in other areas, such as pragmatics (Romero-Trillo, 2008), is much more limited because of the difficulties that such an annotation has, especially in an automatic way and with the existing conflict with the context (Rühlemann, 2011).

Dependency Direction in Liu (2010)

The one we consider the initial and revolutionary contribution of the productions that will come after is Liu (2010). In his paper, we observe a study that tries to reach a typological knowledge similar to that exhibited by Greenberg but through another technique and theoretical approach.

About the methodology, Liu (2010) rescues the approach of Tesnière (1959). This means that syntactic relations are understood in terms of dependency, where there is always a head element and another that depends on it. He proposes a classification of head-initial (centrifugal in Tesnière's terms) and head-final (centripetal in Tesnière's terms) languages, to which he grants a series of typological characteristics that should be better outlined. An example of a head-initial would be "mesa marrón" in Spanish, while a head-final would be "brown table" in English. The study

is carried out by analyzing corpora from 20 different languages (all from Eurasia). One of the differentiating factors is that all the corpora present the ConLL-X tagging form. In other words, although total uniqueness among the corpora is not guaranteed due to the lack of uniformity in the Universal tags that are being developed at the same time as this paper, it is located in a middle path, in which all the corpora offer an identical organization of the information, a fact that facilitates most generalizations.

The research of Liu (2010) allows us to establish a very reasonable typological characterization, considering different types. The dependencies analyzed are mainly verb-subject, verb-object, and noun-adjective. Thanks to their quantification, we find an overcoming of the traditional labels in Greenberg and a greater precision (although the data are conditioned).

The constant concern of how data can be conditioned by existing limitations in Corpus Linguistics (and most especially in those foundational investigations) is something that Liu (2010:18) is very much aware of. He perfectly summarizes and enumerates the existing risks:

(1) How large a treebank is needed in order to get reliable results? How does the size of the treebank influence the results? (2) What influence has the homogeneity of the corpus? (3) What influence do annotation schemes have on the results? In order to find the interrelationship among these factors, Liu (2008) investigated five Chinese treebanks of various sizes, annotation schemes and genres. The results show that these factors may influence the result, but they cannot change the conclusions.

Therefore, as can be seen, on a purely quantitative level, precise modifications of those data can be seen. However, the final result or the linguistic knowledge behind those data is not affected, according to Liu (2008). Having solved the limitations that might have concerned the Chinese researcher, he decides to list the clear advantages of his approach, a fact that catapults the study in this way: “(1)it is statistical and corpus-based; (2)it is robust and non-discrete; (3)it is more fine-grained; (4)it is a more overall typological measure for a language; (5)it can share language resources with Computational Linguistics” (Liu, 2010:21).

A Common Framework

Once the exciting results that could be obtained thanks to this approach were verified, it was decided to finalize a common framework that would allow the most Universal annotation possible for any language. There were different proposals for this. However, the one that has turned out to be the most convenient is, precisely, one that takes the best of different authors (Croft, Nordquist, Looney, & Regan, 2017; Nivre et al., 2023). In this case, we are talking about an annotation scheme that considers the annotation guidelines proposed by Zeman (2008). The morphological tags used

are those proposed by Google (Petrov, Das, & McDonald, 2012). The syntactic reading assumed is by dependencies and fixed with the criteria and tags of De Marneffe et al. (2013); De Marneffe and Manning (2008); De Marneffe et al. (2014). In this case, we are talking about a base in the philosophy of Tesnière (2015), with a set of tags intended to be applicable in any language of the world. The choice for this annotation scheme is due to its greater simplicity from a linguistic and computational point of view. For example, the editors of Tesnière (2015:lxviii) demonstrate how far fewer steps and levels are required compared to a constituent tree analysis, which facilitates better results:

In the 2000s, dependency parsing has boomed, becoming more popular than constituency-based parsing. Among the most popular tools are the MaltParser (Nivre et al. 2007) and MATE (Bohnet 2010), as well as constituency-based parsers that convert the parse into a dependency tree, like the Stanford parser (De Marneffe et al. 2006). Several techniques for dependency parsing are presented in Kübler et al. (2009).

Another of the most commented reasons for the success of dependencies in Computational Linguistics is pointed out by Kahane, Peng, and Gerdes (2023:2): “There is no *a priori* assumption on the locality of a relation, which makes dependency treebanks the obvious choice when attempting to measure tendencies in word order across languages”. These words are similar to those of Kirbi and Gooyabadi (2021), who point out that the best way to get away from any existing bias is to go straight to the data in the crudest and least processed way possible and then formulate theories according to what they suggest.

Despite this growing interest and Liu’s outstanding contribution, Futrell, Mahowald, and Gibson (2015) point out that there are still not too many “large-scale” works to quantitatively compare the properties of languages (they consider 34 languages as “large-scale”). Furthermore, it is true since it is mainly from this date in 2015 that we will start to see different studies in this sense. The paper by Futrell et al. (2015) itself is a clear example. Moreover, they comment, for example, that their study is the pioneer in pointing out the well-known hypothesis that languages with greater word order freedom have more cases in a quantitative and corpus-based way.

This 2015 date is also important because it is also when we already have a mature and fixed version of the unified and Universal annotation scheme of Universal Dependencies (Nivre, 2015; Nivre et al., 2016, 2020). This result, different tags corresponding to different Universal parts of speech, for example, can be seen later, as they will be used in this PhD thesis. This annotation system will also be applied to other tools within the same environment, such as UDPipe (Straka, Hajič, & Straková, 2016), for example, which is an automatic parser of any input in the languages known in UD through this annotation scheme.

Alternative Annotation Schemes

However, the entire scientific community has not unanimously accepted the choice of tags (especially syntactically). Gerdes and Kahane (2016) criticize that there is a mixture of syntactic and semantic criteria that are difficult to assume when carrying out particular research in Linguistics (for example, with the designation of prepositions as noun-dependent), something also mentioned by Croft et al. (2017). It also reflects problems with using specific labels they claim do not reflect orderly linguistic knowledge (e.g., the different ways of marking a modification) (Gerdes, Guillaume, Kahane, & Perrier, 2019).

Therefore, a group of French computational linguists has decided to propose a slight modification of the original UD annotation scheme, in which these previous complaints are taken into account (Gerdes, Guillaume, Kahane, & Perrier, 2018; Gerdes, Guillaume, et al., 2019; Kahane, Caron, Strickland, & Gerdes, 2021; Kahane, Vanhove, et al., 2021). In this case, the use of purely syntactic criteria translates, for example, into a greater weight of everything functional to the detriment of what provides more semantic value. To the basis created by Tesnière (1959) are added the updates proposed by Mel'cuk (1988) from the theoretical point of view. Futrell et al. (2020), for example, are currently starting to use this annotation scheme for their research for the reasons presented here.

They have carried out a conversion of all the corpora annotated in UD to this annotation scheme. However, not many studies in this sense are seen yet, fundamentally, because of the novelty of the proposal and the weight that the original proposal has in this short tradition. The arguments for using it are found in Gerdes et al. (2022). In the authors' words, in UD, there is a loss of syntactic information due to the influence of semantics. This is because there is an unresolved problem in Tesnière (1959) and that De Marneffe, Manning, Nivre, and Zeman (2021) also rescue linguistic heads with a structural center (auxiliary or function word) and a semantic center (lexical or content word). This duality, known as the dissociated nucleus, is the one that generates the problem that SUD tries to solve by granting purely syntactic criteria. The additional features appear more ordered, thanks to the ":" symbol, which makes it possible to hierarchize the labels much more and reduce the multiplicity of single labels. If one does not need or want to use this extra information, one can carry out more general searches without so much specificity, which is impossible in UD.

Another proposal is also observed in parallel called Enhanced Universal Dependencies (Schuster & Manning, 2016). In this case, it is an addition of information not present in UD. No relevant

studies are observed considering this new scheme, and it stands out more from a theoretical point of view. It mainly supports going in the opposite direction of SUD, i.e., taking semantic interpretation to the highest possible level as an annotation criterion (Droganova & Zeman, 2019).

In any case, the Universal Dependencies database proves to be fundamental for typological studies. Its continuous growth and its eagerness to represent as many languages as possible homogeneously allows most of the studies seen here, whether they are more or less closely related to Greenberg. The type of texts used is varied. According to Futrell et al. (2020:384):

The genres of the underlying texts vary from language to language, but most languages have data predominantly from newspapers, blogs, fiction and nonfiction literature, and Wikipedia. Some corpora also include spoken language. Some corpora are of classical, literary, and liturgical languages, including Latin, Ancient Greek, Old Church Slavonic, and Gothic. In the case of Latin and Ancient Greek, the corpora include a great deal of metrical poetry. The corpora of Old Church Slavonic and Gothic consist of religious texts.

For Alzetta, Dell’Orletta, Montemagni, and Venturi (2019:222), this is a problem always pointed out by classical typologists:

In principle, however, differences detected by comparing the distribution of DRs in the ranked list of relations of different treebanks could also originate from genre-specific, rather than language-specific, peculiarities. This is one of the major disadvantages seen by typologists as connected with the use of corpus evidence in typological studies.

However, it should not be forgotten that merely presenting a quantitative summary of structures has no inherent value in Linguistics. Moreover, one must understand these frequencies as alert points of grammatical facts that require a linguistic explanation (Merlo, 2016). We take up this reflection by Merlo insofar as we observe some solely quantitative proposals (fundamentally carried out by computer scientists) in which there is no link to Linguistics or a linguistic motivation, implication, or justification.

New Metrics and Stronger Results

Once we have clear and fixed annotation schemes and adequate databases, we see how interest in this new approach is growing even more in the literature. We find both publications that begin to offer good quantitative studies in Typology and papers that analyze the consequences of this revolution in an even more straightforward way.

An example of the latter could be the paper by Chen and Gerdes (2017:54). This paper contains the main idea of why it is necessary to carry out studies of this type, which allows us to get closer to reality and not to simplify it:

However, just as individual constructions can display varying degrees of syntheticity and analyticality (Ledgeway 2011), different syntactic orders can also be found in the very same language. Reality seems to be messier than we would like it to be. Therefore, probabilities or quantitative approaches, which allow gradual transitions and blurred borderlines, could make some unique contributions on this matter (Liu & Xu, 2012). Moreover, empirical studies based on authentic language data can bring richer details, and then corroborate or improve our knowledge of language classification.

The authors also recognize that the real starting point of this approach is found in Liu (2010). However, they comment that one cannot speak of valuable data for Typology until the birth of Universal Dependencies. This is obviously due to the common dependency annotation framework that is present.

However, this paper also shows very accurate and essential criticisms of the UD resource. It should be noted that many of the corpora hosted in the system have a varied origin and that their conversion to the UD framework has been done later and automatically. This can cause several inconsistencies or unforeseen errors that the linguist must consider when determining the claims. In this case, we believe an excellent linguistic reading of the data obtained already guarantees sufficient reliability for any claim. In the case of research without any linguistic reference, we believe these errors may go unnoticed.

In the case of advantages, the quantitative approach to corpora analyzed by dependencies also allows for obtaining new metrics that allow a richer study. This is the case of Futrell et al. (2020). If Greenberg focused on the direction of the elements, in this case, not only the direction but also the distance, i.e., dependency length, would be considered. It is possible to carry out averages related to the location of two given elements in a sentence. The contiguity or non-contiguity of these elements in different languages may show patterns that have not been well-known until now.

The final result of the research of Futrell et al. (2020) is confirmation of the dependency-length minimization hypothesis. This means that language speakers always prefer a word order that minimizes the total distance of dependencies per sentence. This means, on the one hand, that languages already tend to place their constituents in this close form. If languages show different ways of expressing an idea, the speaker will naturally and unmarkedly choose the one with the smallest distance, comparable to a lower processing cost. This law, which is by no means a novelty, has now been tested in a detailed, quantitative way in many languages simultaneously. In this case, they chose the word (in a graphical sense) to count the distance. There are specific languages in which this would be a problem, but given the novelty of the subject, the contribution is very interesting and relevant.

The most important conclusion of the study by Futrell et al. (2020:391) can be summarized in the quote:

The narrowest interpretation of our result is the following: whatever constraints exist on word order across languages. They have the effect of reducing dependency length beyond what would be expected from projectivity, consistency in head direction, and fixedness in word order. The most straightforward constraint that would accomplish this would be a direct constraint on dependency length itself. However, it is also possible that the observed DLM effect emerges as a side-effect of some other, not-yet-explored factors. The way forward in future work is to implement more and more of these proposed other constraints on languages as baselines and determine whether they can account for the empirical distribution of word orders.

Similarly, we find the proposal of Schnell and Schiborr (2022). It is based on the results of Futrell et al. (2020) to generate knowledge in comparing oral and written corpora. The length of the dependencies in UD is more prominent than in Multicast, which points out something quite logical and has been discussed before. However, in this case, we obtain quantitative evidence in multiple languages. Another exciting aspect of the paper is the approach to Dependency Length Minimization in Typology and corpus, comparing results in different languages and concluding that the principle of short before long is usually fulfilled (when there is such a possibility).

The consequences of this preference for minimizing dependency length do not stop there. The harmonic orders have a very close relationship with avoiding crossing dependencies, which is also related to the ease of processing (Schnell et al., 2022). However, the authors warn that the results are still very much conditioned by the textual Typology and that more work needs to be done on a better balance and a more variable and extensive representation. In addition to the above, Schnell et al. (2022) propose new avenues of research such as corpora of first language acquisition or the mixture of prosody with morpho-syntax (Haig, Schnell, & Seifart, 2022) to obtain new results and confirm the results obtained more resoundingly.

Computational Universals without Linguistics

Not all the publications of these years present this very appropriate balance that we have seen in Futrell et al. (2020) or Schnell and Schiborr (2022). We find studies with a high weight of the statistical and quantitative methods such as Tanaka-Ishii (2021) or Pericliev (2018) where the typological study is merely stated or defined, without going into the existing problems in the discipline and without knowing the problems behind the data shown.

In the case of Tanaka-Ishii (2021) or Pericliev (2018), the existing risk can be summarized by rescuing a quote from Bickel (2017b:118) about the lack of theoretical consistency in this type

of study: “Also, many hypotheses may turn out to be spurious correlations (like the infamous correlation between the declining number of pirates in the world and global warming), because obviously, any statistical pattern is only as good as the explanatory theory behind it”.

Therefore, in this thesis, we advocate a middle way: try to learn about and take advantage of the newest technological innovations without neglecting a search for linguistic knowledge derived from the obtained data. A perfect example of this would be Guzmán Naranjo and Becker (2018). In that case, an analysis is carried out with the treebanks available in that year (quite a few less than now) of some basic directions of some known linguistic types: object direction in languages, subject direction in languages, and so on. This exercise makes it possible, ultimately, to provide a quantitative and text-based proof of an earlier author’s claim based on grammars. In other words, the paper by Guzmán Naranjo and Becker (2018) demonstrates something already known: the non-correlation of subject direction and its linkage to the informational structure of sentences. Now, we have confirmation through this precious data, and we can also see to what degree it holds true in the world’s different languages.

Token-based Typology or Typometrics

In these same years, we observed a high interest from Natalia Levshina in this topic. This can be seen exemplified in the papers Levshina (2019), Levshina et al. (2021), and Levshina (2022). In the first of them (Levshina, 2019), a good review of the state of the art of the subject we are dealing with is observed. It is already interesting that the term “token-based Typology” is fixed, which brings together what we are talking about very well. This paper discusses the benefits of this type of approach’s benefits and the unresolved challenges. Subsequently, the main contributions on this subject, many already presented in this section, are discussed. Then, the new variables that can be obtained thanks to the availability of these data, such as the Shannon entropy example, which calculates the level of flexibility of a language or a specific dependency, are presented. Finally, the paper develops a couple of examples of calculating this entropy metric.

The second paper we cite is very similar (Levshina et al., 2021), as it also shows the benefits and limitations of a corpus-based Typology. In this case, much emphasis is placed on the idea that the word order cannot be reduced to discrete categories and is somewhat gradient. This conception allows a more natural approach to reality and establishes new relationships. The other aspect emphasized in this paper is the consequences of this type of vision and the obtaining of fine-grained data. In other words, the functional implications of this discovery are reviewed, revealing

the motivations that create these patterns: ease of processing, better conceptual accessibility, dependency length, language contact, language change and evolution, vernacular varieties,...

In the last cited paper (Levshina, 2022), a very interesting overview is given about the different possibilities of sources that exist today in corpus Typology. Parallel corpora are available, primarily based on translations of very specific works (New Testament, for example). There is also greater variety in translating different documentation or news from the European Parliament. In the first case, the overly marked topic matter is limited. In the second case, there is the limitation that we only have languages from the same macro-area and with a vast imbalance of Indo-European.

Then, we find the option of arranging comparable corpora. In this case, the texts are not the same but of the same type or subject. In this case, the most recognizable proposal is MultiCAST. According to Levshina herself, there is a significant limitation with the data presented in this case. The languages present are also scarce and do not allow a reliable typological analysis (11 Indo-European languages and a few Oceanic languages). At the end of 2022, we will find a new release: DoReCo. In this case, there are 50 languages (with a particular interest in the typological balance in the selection), and different oral narratives are labeled. The resource's most exciting feature is the phoneme level's alignment. This advantage has a drawback at the morphosyntactic level, as not all this information is available in all corpora.

Finally, Levshina (2022) collects corpora with a unified annotation. Universal Dependencies, the only major project at this level, is presented here. The texts are diverse (positive or negative depending on the type of study to be carried out). The annotation, in this case, is rich and detailed. This project is the one that is growing the most, moreover, because of its open and collaborative science nature.

In the rest of the paper, Levshina shows what correlations can be extracted in (Universal) Typology thanks to these new techniques. She only mentions the papers by Guzmán Naranjo and Becker (2018) and Gerdes, Kahane, and Chen (2019). She certifies that there is much room for improvement in this field. Subsequently, she focuses on exposing the new metrics that can be obtained thanks to this approach, which we have already seen in his previous papers.

This paper mentioned by Levshina, (Gerdes, Kahane, & Chen, 2019), focuses on reviewing Greenberg's Universals 19 and 25. This paper concludes that these two Universals are confirmed thanks to an analysis in 70 languages across corpora in UD. However, we also appreciate how there is a good study on the implications of Greenberg's Universals if they are reviewed from a quantitative point of view. Although in this thesis, we do not share some of the solutions proposed

by these authors; we do consider their reading of the problems of translating Greenberg’s claims to this new methodology to be essential (Gerdes, Kahane, & Chen, 2019:129):

We can remark that Greenberg’s statement is not totally clear. What does it mean that “the pronominal object follows the verb”? Does it mean that pronominal objects always follow the verb, or does it mean that, in most cases, they follow the verb? Is there any quantitative statement hidden in Greenberg’s statement? Whatever the answer to these questions might be, we can translate the statements of Universal 25 into more satisfying, quantitative statements and see whether the implication is verified by our data. In other words, “the pronominal object follows the verb” (V_{pronO}) can be interpreted as: “the percentage of pronominal object on the right of the verb is greater than a ”, where a is some relevant threshold. For instance, for $a = 75\%$, we verify what is a first tentative quantitative Universal: Universal 25: For every language, if the percentage of pronominal objects on the right of the verb is greater than 75%, so is the percentage of nominal objects on the right of the verb. We abbreviate Universal 25 by: $V_{pronO} \geq 75\% \Rightarrow V_{nomO} \geq 75\%$. Universal 25 is illustrated by Figure 5a. Let us recall that the negation of a property $A \Rightarrow B$ is $A \wedge \neg B$. Thus, Universal 25 claims that there is no language with $V_{pronO} \geq 75\% \wedge V_{nomO} < 75\%$, that is, that the corresponding rectangle in Figure 5a (hatched in gray) is empty of any language.

The only solution they propose is a modification of the wording of the Greenberg Universals so that they no longer reflect categories and instead show equilibrium in higher quantitative proportions, as Gerdes, Kahane, and Chen (2019:130) discuss in the quote:

other direction is not to consider a particular threshold at all. For our example, we do not need to propose a threshold, because for almost all languages, we have $V_{pronO} > a \Rightarrow V_{nomO} > a$ for every a , which is equivalent to $V_{nomO} > V_{pronO}$, which gives us the following Universal: Universal 25: Almost every language has a higher proportion of nominal objects than of pronominal objects on the right of the verb. This last statement is verified on our data and corresponds to a near empty triangular form. Universal 25 has no equivalent in terms of qualitative Universals ‘à la Greenberg. Thus working with quantitative data opens up the door to completely new Universals.

The problem we see with this solution is that we are valuing these Universals differently. Therefore, we cannot speak of revising Greenberg’s Universals but of transforming them. This paper shows the germ of what will be seen a couple of years later in Gerdes et al. (2021), which they name as a new discipline, Typometrics.

In Gerdes et al. (2021), we find the presentation of a new tool that allows us to obtain a graphical representation of different morphological parameters with data extracted from UD languages directly. The data presented correspond to basic types. In other words, capturing complex structures or several structures simultaneously is impossible, which only allows the study of elements present in specific and simple Universals. The paper returns to the Universal 25 and shows how it can be converted into quantitative terms and then how it can be represented. It must be remembered that, in this case, we are not carrying out a verification of the Greenberg Universal in a strict way. However, we are using it for quantitative verifications directly.

We have begun to find more proposals about the more direct work between original Universals and this new methodology. The interest in the object of study is seen in multiple conference proceedings such as The Association for Computational Linguistics (2019). The contribution of Morteza, Seweryn, and Ludivine (2019) we do not consider insofar as, although it proposes to do just that, it fails to arrive at such results. In contrast, the proposal of Choi (2021); Choi, Guillaume, and Fort (2021); Choi, Guillaume, Fort, and Perrier (2021) is much more interesting and fruitful. In these cases, we find different papers relevant to the Typology. In the paper, Choi, Guillaume, Fort, and Perrier (2021) work is done to capture the basic word order in Greenberg’s sense of the languages available in UD 2.7. It is not proposed to check Universal 1 directly, which is directly related to this characterization. It is simply the previous step. In Choi (2021); Choi, Guillaume, and Fort (2021), Universal 1 is analyzed with Universals 3, 4, and 17. The contribution of the explicit papers is very interesting and important for this thesis for two reasons, basically, one of them positive and the other negative.

As for the positive contribution of Choi (2021); Choi, Guillaume, and Fort (2021) we must highlight the good methodological work that is carried out, pointing out some of the problems that arise from the attempt to revise Universals made with categories and based on grammars by approaching them with corpora and quantitatively. The limitations or problematic steps in such an operation are very well presented, something that has helped this doctoral thesis.

As for the negative contribution of Choi (2021); Choi, Guillaume, and Fort (2021) we must highlight some essential methodological errors in their research, which make it necessary to question the results presented in each Universal. No one (up to our knowledge) has detected serious errors in the formalizations presented in these works, a fact that generates results that are not representative of the basic word order of each language. Since all the Universals that work are based on this prior implication of basic word order proposed by Greenberg, we believe that we cannot rely on the results presented. As an example, we highlight the most obvious error in our opinion. In the formalization to fix this basic word order, the label “O” (which will represent the direct objects) is assigned a series of limitations on the Universal part of speech that can occupy any element in the direct object dependency. In this case, the parts of speech that can occupy such a position are: “*PROPN|PRON|NOUN*”. That is, proper nouns, pronouns, and common nouns. This contradicts Greenberg’s original formulation, which clearly and explicitly stated that only nominal subjects and objects should be considered (ruling out pronouns). This formalization means that certain structures are captured, which, once they are quantitatively taken into account with the fixed criteria, Arabic, Dutch, German, or Icelandic, for example, are considered NDO. Instead,

based on our formalization data, we propose that the first language be considered VSO and the rest SVO (Brosa-Rodríguez & Jiménez-López, 2023).

In the case of this error applied to subjects, we do not see that it has too great an impact since the tendency of languages to place the subject in the initial position means that both nominal and pronominal subjects do not modify their position. In the case of pronouns, we see a widespread tendency for many SVO languages to shift to exhibit SOV order when this object is pronominal. Moreover, Greenberg was only interested in capturing the first order, which he proposes as a typological predictor because it is the basic one.

We believe this type of confusion in Choi (2021); Choi, Guillaume, and Fort (2021) is due to a lack of clarity on the part of the original UD annotation coupled with linguistic ignorance. The problem with UD is that it distinguishes between two types of subjects: “nsubj” and “csubj”, corresponding to nominal subjects and clausal subjects. Within nominal subjects, there are also pronouns insofar as they are understood as opposed to clausal. When Greenberg asks only to consider nominal subjects, he is not referring to this opposition but is opposing nominal to pronominal. This is another argument for using a cleaner and less confusing annotation scheme like SUD, where everything is “subj”. The lack of linguistic reflection, we believe, leads the authors to explicitly ask for pronouns to be captured as nominal subjects or objects. In short, we do not consider the revisions made by Choi et al., and therefore, we discard their linguistic characterization; we cannot take advantage of it and must carry out our research from scratch, proposing our basic word order to revise Greenberg’s Universals.

At the end of 2023, we found a paper by the proponent of this approach, Haitao Liu (Yan & Liu, 2023). This shows that the researcher still believes in the potential of this methodology and the relevance of the object of study. This paper proposes to extract interesting metrics that allow a statistical formulation to examine the first five Greenberg Universals. UD version 2.5 is used, and only 74 corpora are reviewed. We are not going to offer a depth of the findings about these Universals, nor are they taken into account in this work, precisely because when we reviewed the supplementary material where the formalizations carried out are shown, we detected the same error we have previously exposed. That is, Yan and Liu (2023) also consider the occurrences of pronouns to determine the basic word order of a language. This choice, once again, generates problems from a typological point of view, in addition to the fact that, in our view, it is impossible to speak of a translation of Greenberg’s proposal to the quantitative because it adulterates his words.

In conclusion, we believe that this closure is rather symbolic. The one who can be considered the father of quantitative Typology through corpora more than a decade ago is still, at the end of 2023, interested in improving the results obtained. Despite the new avenues of research opening up and the new possible metrics, some aspects can be improved in such publications that allow a line of improvement. From this approach, the study of Greenberg Universals interests many researchers, but we always see the study of the same Universals (those with a simpler structure). This fact means that there are Universals that have not yet been reviewed in this way, and, therefore, new things can be proposed (in addition to proposing rectifications or improvements for Universals already formulated in an imprecise way, in our opinion). Moreover, no comprehensive study of Greenberg’s Universals, i.e., one that tries to understand the limits or ways in which the totality of Greenberg’s Universals can be analyzed, has been observed either. Hence, the interest we have in this doctoral thesis: we will try to answer whether each of the Universals proposed by Greenberg is treated (or not). This fact, *a priori*, will facilitate a greater and more complete knowledge of the American author’s work.

10.3 Typology and Universals

In the previous sections, we have been closely relating a conception of Universals (the one analyzed in this thesis) to Linguistic Typology, a term already appearing in von der Gabelentz (1869, 1891, 1894). There is no single, settled conception of the similarities between Universals and Typology and their kind of relationship. One of the main problems is that we find different visions and definitions of what Typology is.

The opposition between Universals and Linguistic Typology

For Hockett (1963:3), the existing relation is one of opposition: “If some feature is indeed Universal, then it is taxonomically irrelevant”. Therefore, he understands Typology as a science that deals with the classification of languages. In this case, the criterion for classification would be linguistic, as opposed to other classifications such as genetic or geographic (Daniel, 2010; Greenberg, 1960). The logic behind the quotation of Hockett (1963) is that if we formulate a rule applicable to 100% of languages, it will not allow us to group languages around such an element since all languages would be in the same group.

This taxonomic view appeared a few years earlier in Jakobson (1958). He compares the classification of languages with the classification that can be made of animals. In such an exemplification,

it is shown how there are criteria that are not useful for the meaningful classification of elements. For example, if we were to distinguish by color, we would not obtain satisfactory results. In short, Jakobson understands Typology as a taxonomy that allows us to offer reliable and meaningful elements. These, we understand, will be the ones to be used later in the Universals.

Universals as the father of modern Linguistic Typology

In the same sense, Cooreman and Goyvaerts (1980) understand Universals (we shall understand if we do not specify the contrary, implicational, statistical, and superficial) as the basis of modern Typology. They believe that the criteria used and the starting point of Typology arise from Greenberg's Universals. In other words, they see Universals as the precursors of Typology as we understand it today. This aspect is something that Comrie (1989) will also emphasize, when he comments that the point of union of Typology and Universals has the basis in the word order implicational Universals formulated by Greenberg.

Precisely, the book that deals with the subject of this section extensively is Comrie (1989), which we recommend to expand on this topic, together with Moravcsik (2012). First, Comrie (1989:33) collects views such as the one shown above in Hockett (1963):

At first sight, the study of language Universals and the study of language Typology might seem to be opposites, even in conflict with one another: language Universals research is concerned with finding those properties that are common to all human languages, whereas in order to Typologize languages, i.e., to assign them to different types, it is necessary that there should be differences among languages. The contrast can thus be summed up as one between the study of similarities across languages and the study of differences among languages.

In such a quote, we can see how he values the possibility, from a rather general point of view, of seeing the relation between Universals and Typology as opposites. He explains that the search for Universals needs to find common features in languages, while for Typology, it is necessary to find differences between languages. He summarizes this as the contrast between the study of language similarities and differences.

Universals and Linguistic Typology are almost the same

However, Comrie (1989:33) referred to a superficial approach to the phenomenon. Subsequently, he comments that:

Nevertheless, in practice, the two studies proceed in parallel: typically, linguists who are interested in language Universals from the viewpoint of work on a wide range of languages are also interested in language Typology, and it is very often difficult to classify a given piece of work in this area as being specifically on language Universals as opposed to language Typology or vice versa.

Then, he states that it is very difficult to isolate the study of Universals from Linguistic Typology since both approaches share the same interest. Although he does not say so, this interest can be perfectly seen in the implicational Universals, which try, at the same time, to look for invariants from a linguistic characteristic and to know better the different groupings that can be made of the languages of the world from a linguistic point of view.

In other words, for Comrie (1989:34), there is only a slight difference: “Language Universals research is concerned primarily with limits on this variation, whereas typological research is concerned more directly with possible variation”. In other words, both Universals and Typology study language variation. Universals deal more directly with the limits of this variation (structures that are not possible given a precondition, for example). Typology focuses more freely on the different variations (linguistic types).

Linguistic type is understood as “una combinazione di proprietà strutturali logicamente indipendenti le une dalle altre, ma reciprocamente correlate”² (Grandi, 2020:11). That is, as a linguistic group based on common linguistic criteria. Alternatively, in the words of Hetterle (2015:9):

A major assumption is that linguistic structure varies in countless ways, but that the variation is highly systematic, reflecting Universal properties of human communication, cognition, and social interaction, as well as principles of language change; that is, it ultimately reflects the nature of human language.

Returning to the words of Comrie (1989), we can understand that if we design all possible combinations of the manifestation of a phenomenon and all the types are represented in different languages, it is of interest for Typology but not for Universals. On the other hand, if there is some combination in which no occurrences of languages of the world are observed, this is of great interest to scholars of Universals.

Grandi (2020:61) speaks of Typology as if it were of Universals; for him, it is almost the same. For example, the chapter entitled “Typology and Syntax” will deal with the content of this thesis on Universals (Greenberg). Moreover, we find explicit references to it as in:

Da sempre i percorsi della tipologia linguistica e della ricerca sugli Universali si intrecciano e si sovrappongono, condividendo, almeno in parte, metodi di indagine e strumenti esplicativi, al punto che oggi pare impossibile trattare dell’una senza fare riferimento all’altra.³

²This could be translated into English as: “a combination of structural properties that are logically independent of each other but mutually related”.

³This could be translated into English as: “The paths of Linguistic Typology and research on Universals have always intertwined and overlapped, sharing, at least in part, methods of investigation and explanatory tools, to the point that today it seems impossible to deal with one without referring to the other”.

In other words, Grandi tells how the paths of Typology and the search for Universals cross and overlap, sharing, at least partially, research methods and explanatory tools to the point that it seems impossible to deal with one without referring to the other.

We also find, as in the case of Comrie, how Grandi (2020:61) mentions a possible superficial view of the relationship between the two concepts as antagonistic elements:

Eppure, a prima vista, la tipologia e la ricerca sugli Universali paiono perseguire obiettivi diametralmente opposti. La prima, si è detto, si occupa della variazione interlinguistica, cioè di come le lingue si differenziano le une dalle altre. La seconda, non c'è bisogno di evidenziarlo, studia ciò che è comune a tutte le lingue, concentrandosi dunque sulle proprietà rispetto alle quali le lingue non possono variare.⁴

However, the contribution that we consider most relevant in Grandi (2020:63) refers to a very obvious aspect but is not made explicit in other publications:

È evidente che gli Universali assoluti non lasciano alcuno spazio alla variabilità: essi consentono di identificare un unico tipo linguistico, cui afferiscono tutte le lingue storico-naturali del passato e del presente. In questo senso, essi hanno davvero poche occasioni di cooperazione con la tipologia.⁵

The above quote clearly shows how the non-implicational Universals (“absolute” ones in Grandi’s quote) do not come into play in this type of equations since by referring only to a single linguistic type that encompasses all the languages of the world (there is no prior restriction), there is little relation with Typology. In other words, whenever we speak of the close relationship between Typology and Universals, we are referring to the implicational Universals, something that all researchers surely take into account but which they do not mention and should (Grandi, 2020:64):

Gli Universali implicazionali pongono in relazione due (o più) proprietà, vincolando la presenza di una di esse alla presenza dell’altra. In altre parole, essi affermano che un tratto linguistico può relizzarsi in una lingua storico-naturale solo se nella medesima lingua è attestato anche un altro tratto linguistico.⁶

Grandi (2020:64) concludes this section by making even more explicit, if possible, what it is that makes the implicational Universals those that have this typological essence:

⁴This could be translated into English as: “Nevertheless, at first sight, Typology and research on Universals appear to pursue diametrically opposed goals. The former, it is said, deals with interlinguistic variation, that is, how languages differ from one another. The latter, it goes without saying, studies what is common to all languages, thus concentrating on the properties in respect of which languages cannot vary”.

⁵This could be translated into English as: “It is evident that absolute Universals leave no room for variability: they allow the identification of a single linguistic type, to which all natural-historical languages of the past and present belong. In this sense, they have very little occasion for cooperation with Typology. on the properties concerning which languages cannot vary”.

⁶This could be translated into English as: “Implicational Universals relate two (or more) properties, binding the presence of one of them to the presence of the other. In other words, they state that a linguistic trait can be realised in a natural-historical language only if another linguistic trait is also attested in the same language”.

Per la tipologia, il contributo degli Universali implicazionali ha un rilievo decisamente maggiore di quello offerto dagli Universali assoluti. Ponendo in relazione due proprietà distinte e teoricamente indipendenti, infatti, un Universale implicazionale lascia alle lingue un buon margine di reciproca differenziazione e, quindi offre parametri affidabili e attendibili per lo studio della variabilità interlinguistica.⁷

That is, Grandi shows how elaborating an implicational Universal further deepens the differentiation of languages, allowing for such cross-linguistic variation simply by offering more data that confirms and expands the difference between types.

Universals and Linguistic Typology are the same

If we have seen considerations of the Typology-Universals relation as contrary and complementary, we can also offer views where they are considered the same. This occurs, for example, with Ramat (2009:1), who defines Linguistic Typology as: “the systematic cross-linguistic comparison that aims to discover the underlying Universal properties of human language”.

Therefore, the Universal is part of the definition of Typology for Ramat (2009). It can be understood that the study of Universals is synonymous with the study of Typology or vice versa.

Universals and Linguistic Typology are the same and the opposite

Although Hockett’s words are difficult to refute, Saporta (1963:48) understands it as: “The two are merely different sides of the same coin”. In such a metaphor, we find again the relation of opposition but, simultaneously, a feeling of undeniable pertinence and interdependence. Both proposals are necessary and useful for the other. A Universal that ensures that languages having SOV order will have postpositions requires prior knowledge of which languages are part of the SOV linguistic type. The fulfillment of such a Universal allows, precisely, to know better the behavior and characteristics of the SOV type and not of the world’s languages in general.

Daniel (2010:1), on the other hand, chooses to present, in the definition of Typology, the relation between Typology and Universals as opposites and the same at the same time:

Linguistic Typology compares languages to learn how different languages are, to see how far these differences may go, and to find out what generalizations can be made regarding cross-linguistic variation. As languages vary at all levels, Linguistic Typology deals with all levels of language structure, including phonology, morphology, syntax, and semantics.

⁷This could be translated into English as: “ For Typology, the contribution of implicational Universals is much more important than that offered by absolute Universals. By linking two distinct and theoretically independent properties, in fact, an implicational Universal leaves languages a good margin of reciprocal differentiation and, therefore, offers reliable and trustworthy parameters for the study of interlinguistic variability”.

The term Typology encompasses, for Daniel, both the search for differences between languages and invariance between them. Following the metaphor used by Saporta (1963), we could say that he is describing Typology as the whole coin, each of the two elements mentioned in the previous sentence being the two sides.

Moravcsik (2012:9) shows this same dual vision:

The field of linguistic research called language Typology is the study of the latter two phenomena: typologically and Universally shared features of languages. It focuses on the concept of a language type. The term “type” in everyday usage is synonymous with “kind”: it refers to a subclass of a class of entities. In this broad sense, two languages belong to the same type if they have at least one characteristic in common, regardless of whether this shared characteristic is due to shared inheritance or borrowing or similar environmental conditions. In actual linguistic usage, however, two languages are generally said to belong to the same type if their similarities hold across various genetic, areal, and cultural groups.

For Moravcsik (2012), Linguistic Typology encompasses both studies focused on looking at the characteristics that languages exhibit to create a type and the formulation of patterns that cut across all languages.

A similar view was presented by Greenberg (1978b:34) when he tries to capture Linguistic Typology as a research method: “By a singular turn of fate, a method which in the past has been used to characterize and define differences among languages has become a major methodological tool in the investigation of linguistic generalizations”. This method was used to characterize the differences between languages and has been used to do just the opposite, to test invariance. That is to say, for Greenberg, the existing relationship is based on the fact that Typology is the methodology used in the study of Universals.

Universals as necessary for the Functional-Typological Approach

Croft (2003:1) also echoes the concept of linguistic type to define what Typology is: “a classification of the phenomenon under study into types, particularly structural types”. As we can see, this definition does not provide any concept not seen before. We highlight this definition precisely because Croft uses the word defined to explain better the two phenomena we are dealing with. In other words, he makes a distinction between “typological classification” and “typological generalization” (Croft, 2003:1). In this use of labels and this confrontation, the cross-linguistic basis of both approaches is seen, even if not made explicit. Such a comparison of labels makes it possible to see the common points, in this case, Typology *per se*. It is the interest in studying the world’s languages and their behavior around linguistic criteria. About the differences, Croft (2003) points out that, in one case, it is only a matter of classifying these languages, that is, of grouping them

around the presence, absence, or particular distribution of a certain element. In the other case, it is a matter of formulating a rule that captures the general behavior of the world's languages.

We must not forget, moreover, something we have already seen but Croft (2003:2) comments about these two concepts:

This view of Typology is closely allied to functionalism, the view that linguistic structure should be explained primarily in terms of linguistic function (the Chomskyan approach is contrastively titled formalism). For this reason, Typology in this sense is often called the (functional-)typological approach. [...] empirical scientific approach.

Both Typology and (typological) Universals share a functionalist view of Linguistics, characterized by an interest in explaining the causes and consequences of the elements under review. Indeed, in Croft (2003), we do not believe that the three elements cited above are simply presented as interpenetrating and differing in some respects. We observe a much more global conception of the phenomenon, with a concern to understand them as a spatial-temporal correlation. In other words, we can understand that we are dealing with the three steps necessary to conduct a linguistic investigation of this genre:

1. **Typological classification:** This allows us to know the linguistic characteristics of the world's languages in a cross-linguistic way and to be able to group them.
2. **Typological generalization:** which is characterized by the desire to understand the behavior of the world's languages around a previously characterized type.
3. **Functional-Typological Approach:** which would be the search for the motivation that has led to the emergence of such a Universal in such a linguistic type.

Such a conception of different steps of a whole contributes to the belief that we are dealing with an empirical science that has been criticized in previous years, according to Croft (2003), since only the first of the steps was considered a Linguistic Typology.

This view, in a different way, can be seen in Bickel (2007). In this reference, Linguistic Typology is initially presented as a mere alternative and competing methodology to generativism. Therefore, they shared the same goal (limits of what is possible in a human language). However, according to Bickel, traces of full independence (an autonomous discipline) have become apparent in recent years, with its agenda, problems, and theories. In this case, he comments, the goal is no longer the limit of what is possible or impossible in languages, but to explain why something occurs or does not occur, something that detail and expand Booij (2009) and Bach and Chao (2009).

Finally, we find the proposal of Mauri (2022b:24), which offers an interesting argument for the point of view of the integration or interrelation of the two concepts:

Andando alla ricerca dei limiti all'apparente caos linguistico del mondo, la tipologia cerca di scoprire gli aspetti comuni, regolari e prevedibili, che caratterizzano in modo Universale le lingue umane, facendo emergere quell'ordine nascosto che permette di spiegare come mai tutte, per quanto diverse, siano ugualmente funzionali al loro scopo.⁸

In the above quote, Mauri comments that Typology aims to discover the common, regular, and predictable aspects that characterize the world's languages universally, making it emerge, within the apparent diversity and chaos of the world's languages, that they are all the same from a functional point of view and behave the same.

If we can determine the existence of some Universals, we assume that languages behave in the same way in certain aspects. Therefore, these Universals can be understood as proof of these language preferences, regardless of their other characteristics (or considering one of them as a trigger, depending on the type of Universal). The reason for explaining the motivation of the Universal is the ultimate purpose of Linguistic Typology (Functional-Typological Approach).

There can be several causes that generate Universals. Typologists always start from a functionalist view of the phenomenon: everything has a functional explanation. The most common case is the one put forward by Campbell et al. (1988:221): “Word-order patterns are arranged to make processing easier”. This view, moreover, accords very well with the naturalization of exceptions. It is possible to find languages that are an exception to the word-order pattern formulated. Instead of considering this Universal as useless or the language as impossible, it is simply considered a more difficult language to process (Moravcsik, 2012).

There are other preferences. If the interest of one coincides with the interest of the other, there is no problem. If they contradict each other, one will be fulfilled by the language, and the other will not. This is what happens, according to Greenberg through Dryer (1988a), with the possibility of finding OV languages with the adjective before or after the noun. There is harmony consistent with the tendency to place the dependent before the head and dominance. The latter would apply more strongly to an adjective because it is a modifier, whereas it is more difficult to see dominance in the case of a direct object since it is an argument. Therefore, Greenberg and Dryer's explanation indicates that harmony dominates in OV languages that prefer to place adjectives before the noun, while in OV languages that place the adjective after the noun, dominance dominates.

⁸This could be translated into English as: “By searching for the limits to the world's apparent linguistic chaos, Typology seeks to discover the common, regular and predictable aspects that Universally characterize human languages, bringing out that hidden order that makes it possible to explain why all of them, however different, are equally functional to their purpose”.

This explanation is, to some extent, unsatisfactory. Dryer himself shows that the direction of the adjective is a non-correlation pair concerning the object. Pointing out these two options, the Universal does not allow prediction of any kind.

These kinds of explanations bring the investigation of Universals in Typology much closer to the more Chomskyan approaches not dealt with in this research (Bever, 2009; A. Clark & Misyak, 2009; Müller, 2009). Comrie (1989:23-24) presents the justifications for the two approaches very well. He compares the justification that Universals exist because they are innate with the justification that “arguing certain Universals serve to make the language more functional, either as a communication system in general or more particularly relative to the communicative needs of humans”. The clear differences in justification are also evident here (as is the different use of methodologies to study Universals). However, the same interest in understanding the implications of the Universal is apparent.

In the same way, Bickel (2017b:115) explains why, on the point at hand, there is a rapprochement between typological and Chomskyan conceptions of Universals, even though they remain distinctly different things:

The relevant factors underlying statistical Universals draw from the same psychological and biological issues that also matter to the theory of absolute Universals, and it is no surprise, therefore, that research on statistical Universals is also labeled “cognitive” or “biological” (e.g., Givón 2001, 2010, Croft 2003). However, the sense is a different one. For statistical Universals, psychological and biological facts define causal theories that predict how languages develop over time. A classic example comes from word-order research: Hawkins (1994, 2004) proposes that the way the brain processes languages causes languages to favor certain structures over others, for example, “harmonic” over “disharmonic” patterns. Harmonic patterns are ones where nested phrase structures are oriented in the same way, e.g., in English, the head (constitutive element) of noun phrases is initial (the development in the development of science) and so is the head of verb phrases; in Turkish or Hindi, both heads are final. This is indeed what we find in most languages, but there are exceptions. Rather than trying to reanalyze the exceptions so as to fit an absolute Universal, in research on statistical Universals, the exceptions are taken at face value, and the relevant Universal is taken to be probabilistic.

Researchers such as Hawkins (2004, 2009) focus on what he names Minimize Domains. According to this proposal, languages try to unite dependent elements as much as possible precisely to facilitate their processing and identification. An example of this, offered by the author, is the possibility of finding, as an unmarked basic word order: “went to the movies”, “the movies to went”, “went the movies to”, and “to the movies went”. According to Hawkins (2009), the first option would be the one observed in English, the second in Japanese, and the other two would hardly be seen in the world languages (VO+postposition or OV+preposition). This is because these configurations place the preposition further away from the verb head. These preferences are embodied in the union of VO+preposition and OV+postposition. Therefore, languages can freely

exhibit different configurations (being OV or VO, for example), but one of these choices conditions many others (the adposition direction).

Subsequently, the same Hawkins (2010) decides to put forward what he considers to be the three basic rules that support the existence of Universals. In addition to the Minimize Domains we have shown above, he also proposes Minimize Forms and Maximize Online Processing. Minimize Forms corresponds to the principle of economy, which governs the simplification of everything as much as possible (regardless of the linguistic level). This is not easy to apply to word order, but it can be seen in different formulations where it is clear that the most frequent is the simplest. For example, in the relationship between word length and word frequency. In the case of Maximize Online Processing, it is the movement of the element to be highlighted which is important for processing. A very clear example can be the typical movement of languages to highlight the required information in the first position in the WH question.

Thanks to the availability of new data through a corpus approach computationally, we have also obtained new metrics indicating other types of preferences for all the world's languages. The clearest example is in Futrell et al. (2020:391). They show that Dependency Length Minimization is Universal and affects all languages of all types equally. It is especially visible in languages with higher flexibility, which allows technically the embodiment of different structures with different orders. Using these languages shows how the preferred option in language use is always the one with the lowest dependency length. This means there is a preference for placing the head and its dependents as close together as possible. This is understood as facilitating processing.

Other authors emphasize the role of diachrony and interlanguage influences as the real generators of exceptions rather than the competition of underlying factors. By this, we mean that a language can change its type from SVO to SOV (by contact with a neighboring language, for example). In Greenberg's Universals, SVO languages tend to be prepositional, and SOV languages tend to be postpositional. If this change occurs, the change in the direction of the adposition may not have taken place, which would generate an exception to this rule. However, if the formulated Universal is true, over time (for harmony and ease of processing), the previously formulated change will encourage this new change. Therefore, the exception in that Universal could be understood as the consequence of a snapshot of a change that is in process.

If, after some time, this exception does not disappear as we have said, and if it is exclusive to a particular language, we then consider it *rarum*, according to Cysouw and Wohlgemuth (2010), Malchukov (2010) or Harris (2010). Therefore, we believe that both the above justification for

the presence of exceptions and the justification that various competing factors are not mutually exclusive and can coexist. In other words, depending on each case and each Universal, motivation can have one weight or another without the need to choose *a priori*. This is also shared by Christiansen and Chater (2009:452), who point out that the “evolution of language does not, of course, take place in a biological or social vacuum, but rather, is shaped by multiple constraints”.

In other words, Bickel (2017b:115) sums it up with the quote: “These historical processes are influenced by a large array of factors, and because these factors are in highly dynamic competition, Universal effects manifest themselves only statistically, never categorically”. These factors that generate exceptions, according to Levshina et al. (2021:14), are, specifically:

Word order change may be caused, first, by language contact (see Section 2.3 and also Bradshaw 1982; Dogruoz & Backus 2007; Heine 2008, etc.), by substrate influence (which could be subsumed under language contact), and by internal change. There is a large role for language- or lineage-specific pathways in word order change (cf. Dunn et al. 2011). Some tendencies may apply to all or most languages. For example, Maurits and Griffiths (2014) propose a (Universal) tendency for languages to change to SVO order; a gradient-based approach could identify through which processes such change proceeds.

From what has been seen above, Universals can be seen as communicative strategies so effective that they are shared by the languages (Grandi, 2020). Instead, Universals are the evidence or realizations that demonstrate such application. According to Grandi (2020), the three leading causes can be summarised as economy, iconicity, and communicative motivation. They can be compared with the abovementioned classification in Hawkins (2010) since they share the same conception. The first one groups together the intention to speed up or optimize communication without losing the communicative potential. The second refers to the tendency to reproduce sequences structurally based on how the information to be transmitted is organized mentally. The last one refers to the intention to clarify and facilitate the process that allows decoding the information in the most effective way possible (Grandi, 2020:67-69).

Whatever the motivation that may trigger a Universal, it is essential that it exists or can be formulated. This is why we believe that the work of the linguist is fundamental. In some purely computational studies, we see analyses that do not have the linguistic object behind them in mind. This, in an extreme case, can lead us to what Sansò (2022) or Tanaka-Ishii (2021) point out about risks in statistics, where it could be assured that there is a clear relationship between two independent variables that are actually not true. In other words, correlation does not always correspond to causation.

Bickel (2017b), in addition to aspects that other authors have already discussed, offers a historicist overview that helps to understand better what is happening with Linguistic Typology studies.

He broadly presents an apparent distinction between the studies of the first half of the 20th century (search for differences between languages) and the second half of the 20th century (search for common patterns in languages). We believe that this overview that he presents can help us better understand the change we have observed in the different definitions and relationships proposed in this section. As we have seen, the first approaches marked the opposition between Universals and Typology more clearly and forcefully. We believe this is due to temporal proximity with the period of rupture of the previously described conception. The need to draw clear boundaries was greater. On the other hand, it can be seen how, little by little, from the 1970s onwards, culminating in our times, a much more globalizing vision is advocated, where the boundaries between these two concepts are blurred or even superimposed.

Alzetta et al. (2019:209), on the other hand, point out that this change occurred at the beginning of the 21st century: “Over the last two decades, Linguistic Typology shifted its focus from the quest for Universals to the study of the distribution of the features underlying linguistic diversity”. However, we do not consider it accurate, as we have seen how, throughout the second half of the 20th century, examples of the feature distributions are seen. They refer to such a change based on Bickel (2015). It would be more accurate to refer to functional explanations more directly, which is what Bickel refers to. In such a case, it is also necessary to clarify that some examples of such interest are already seen in the 20th century.

Finally, in the most recent years, we have not found any difference or novelty concerning previous visions. It could be said, in general, that we continue to find the vision of seeing Typology and Universals as synonyms.

Regardless of the reason that best suits each case, we have seen an eagerness to relate the appearance of Universals to any kind of benefit in human communication. Therefore, it is important not to understand Universals as an end in themselves but as a means to gain knowledge of the functioning of languages and their preferences, whatever they may be.

UNIVERSITAT ROVIRA I VIRGILI

GLU - A GRADIENT MODEL FOR LANGUAGE UNIVERSALS. A COMPUTATIONAL REVISITATION OF GREENBERG UNIVERSALS

Antoni Brosa Rodríguez

11

Limitations in Typological Universals Research

Contents

11.1	Unawareness of Total Languages of the World	138
11.2	Dialects and Languages Distinction	139
11.3	Dead Languages and Creoles	140
11.4	Scarcity of Data	142
11.5	Anglocentrism	144
11.6	Sampling	148

Another central aspect to be mentioned when investigating Universals in language is the limitations that exist in their study and that must be taken into account. Some solutions have been offered; in other cases, it is difficult to do anything. For Perkins (1989:297):

We are forced to make inferences about language without looking at all possible cases, i.e. by using samples of languages; this is necessary due to limitations of time, money, existence (some languages no longer exist and others have not yet developed), and descriptions (only some of the world’s languages have been studied and described).

We summarise the linguistic or extralinguistic issues to be taken into account with the sections:

1. **Unawareness of total languages of the world.** We do not know the exact number of languages worldwide; our knowledge is incomplete. Some aspects of the other points condition this number.
2. **Dialects and Languages distinction.** There is no unanimity on distinguishing some varieties as languages or dialects.

3. **Dead Languages and Creoles.** Some authors consider that varieties such as Creoles and dead languages should be considered in this total number and the analyses. Others do not.
4. **Scarcity of Data.** Even if we knew the total number of languages worldwide, the data available on most of them (for typological research) would still be very limited.
5. **Anglocentrism.** The weight of European languages, mainly English, can lead to misunderstandings about how some languages work.
6. **Sampling.** The solution for a correct typological study is an adequate selection of the languages analyzed that represent the total number of languages. However, there are different opinions about sampling.

11.1 Unawareness of Total Languages of the World

One only has to look at the figures from Typology experts on the world's languages to see no consensus. In addition to the constant evolution of this number, there is no consensual agreement on the criteria to be followed. The lack of knowledge of some of the world's languages in certain areas (such as Austronesia) with hundreds of them does not allow for a precise categorization shared by all linguists. There is a difference of more than twice as many languages between some proposals in the literature.

For Comrie (1989:9), this number is around 4000 languages:

the estimated number of languages spoken in the world today is so large that, if we were to await investigation of each languages before embarking on research on Universals, we would probably never even initiate our main task. Although estimates of the number of different languages vary considerably, they tend to cluster around the 4,000 mark.

This number goes up to 5273 languages for Ruhlen (1987), who decides to sort around 20 groups or large families. Rijkhoff and Bakker (1998) also shows disagreement regarding the different families, with different authors going from 27 to 120. Ponti et al. (2019) prefer to go up to the 141 groups offered by the Ethnologue (Eberhard, Simons, & Fennig, 2023), which currently sets the languages present at 7168 in 2023, a number lower than that of Glottolog (Hammarström et al., 2023): 8595.

Hammarstrom, in private conversations, offers the count of 6531, which will be published in a future paper. This is in between the generalization of Croft (2003) and Evans (2010) (6000 languages) and that of Pagel (2009), Grimes (1988) and Ruhlen (1991) (7000 languages).

However, other aspects also influence and make the study complex. We also want to maintain the comparison made above between languages and biology to extend it to the existing problem with species in the world: species that we do not know yet, species that die, new species that emerge, species of which we only know the name but know very little about them, two subjects that we may believe to be of the same species but are not, etc.

We believe that giving an exact figure is misleading and fictional for all these reasons. We prefer to assume this limitation naturally and think we are talking about a figure of around 7000 languages. In no case do we come to believe that the problem is that we are dealing with an opinion rather than a science, as claimed by Anderson (2005).

11.2 Dialects and Languages Distinction

One apparent cause of the identification of the total number of languages in the world is that we consider that we have not yet resolved or reached a consensus on what a language (as opposed to a dialect) is. This issue has been controversial since its inception. It can be seen, for example, clearly in the well-known aphorism attributed to Weinreich, which defines a language as a dialect with army and navy. Therefore, we must recognize extralinguistic (not to say directly political) criteria as the cause of these problems. It is what causes us to gather different varieties under the label of Chinese that are not mutually understood by the supposed speakers of that single language, for example.

However, Linguistics has tried to offer objective and precise linguistic tools to carry out a less arbitrary and biased characterization. The criterion that has been most used in these cases is that of mutual intelligibility: “Languages are not clearly definable objects. Mutual intelligibility is the commonly accepted delimiter used to distinguish languages from dialects” (Perkins, 1989:295).

However, this criterion often remains insufficient. Not only because of the substantial weight of extralinguistic criteria that overlap with linguistic ones but also because there may be problems not foreseen in a superficial approach to the problem. The delimitations may not be entirely transparent and sharp: they may occur much more in one direction than in another, at some levels than at others, they cannot be quantified, etc. Thus, Hammarström (2008:34) clearly explains that: “mutual intelligibility is not a strict yes/no distinction but a matter of degree”.

We recommend extending and reviewing all the problematic cases raised in Hammarström (2008), which cause this series of problems when assigning a number to the world’s languages. For

example, Hammarstrom points out the case where a language A has mutual intelligibility with B, B, and C, but A and C do not. It must be well justified if it is considered that we are dealing with the same language, 2 or 3.

Lahiri and Plank (2009) also assume this problem but downplay its importance: if something is Universal in languages, it will also be Universal in dialects (and other varieties). However, we believe that this mismatch has an impact on sampling. Since we cannot sample the totality of languages, we must make a selection, as we shall see. This selection cannot be unbalanced by this issue and must be considered. This is especially relevant in typological Universals (which try to describe different types). The dialect-dependent characterization of a language is something that must be taken into account. For example, if we rescue the study of Corbett (2009), even in English, we find this problem. The word “committee” may show a different number value depending on the variety of English used. These aspects must be considered when giving only one label as representative of a language (or the opposite in the case of languages understood as dialects).

In recent years, we have found some corpus-based models that attempt to determine this distinction more objectively and cross-sectionally (Schnell & Schiborr, 2022). However, we do not consider the results obtained so far particularly relevant, as the current data limitations do not allow for an accurate and faithful representation of languages and, especially, dialects.

11.3 Dead Languages and Creoles

As Hockett (1963:2) says: “If we had full information on all languages spoken, there would remain languages recently extinct on which the information was inadequate”. This means that even if we were to solve the above problem (although we cannot), we would still not have the exact number of languages in the world. The number is changing (increasing or decreasing) due to the birth and death of different languages. Concerning dead languages, we believe that it is not only, as Hockett says, that we have inadequate information. We also believe that, in some cases, we have no record of languages that existed and died. However, in cases where we have limited data on them, we also have to be aware that the evidence that has come down to us may not be entirely representative, as Magni (2009) points out about the poetry of these languages. Bakker (2010) puts the number of languages that have existed at about 240,000, and he claims that we are missing a great deal of information. Pagel (2009) further explains that 10000 years ago was the time of most remarkable linguistic diversity. However, we believe these laments cannot go beyond this point since there is

no possible solution. Lahiri and Plank (2009:33) sum it up very well in the quote:

The empirical basis of Universals research can only be (a sample of) a subset of the domain for which Universals could maximally claim validity, and have traditionally been claiming validity: that of humanly possible languages. Therefore, the only viable domain for Universals research, then, is all-languages-present-and-past-as-known-to-us-now. The risk, inevitably, is that this subset is atypical of the set of all-languages-ever: what appear to be “Universals” may, *sub specie aeternitatis*, only be historical contingencies.

For all these reasons, we understand that very few things can be done with disappearing languages that we do not know and will not know. However, we believe it is necessary to mention this situation and make it explicit in the studies it influences to be aware of it. Moreover, it is important because this is an extra reason to be firm with language preservation (Schnell et al., 2022).

Comrie (1989), in addition to past languages, talks about future languages. He proposes incorporating dead languages if there is sufficient data on these, but it is unnecessary to be obsessed with this fact. Croft (2003) speaks of the linguistic uniformitarianism that governs the studies, according to which it is presumed that the rules that govern the current languages are the same rules that govern the languages that are no longer there and those that will be. Therefore, in the variety that we find in today’s languages, all the possibilities that existed in the past in dead languages can be gathered, even though it may be believed that some particular linguistic type has disappeared.

Dryer (1989:9) also agrees with this vision that allows us to claim Universals without being invalidated by dead languages:

At least within a time-span of several thousand years in either direction from the present, there has been no significant sense in which human language has evolved; i.e. no sense in which human language as a whole today is different in essence from that on ten thousand years ago; more specifically, it assumes that all human languages spoken today represent the same level of evolution.

In other words: “the range of human languages spoken in the world today is sufficiently large and varied to include examples of virtually all the kinds of structure that are possible in human language” (Dryer, 1989:9).

Therefore, the interest in Typology for Universal patterns in the changes (Greenberg, 1978a; Greenberg et al., 1963) allows us to understand Universals from a diachronic perspective. Indeed, we must remember that for many researchers, exceptions occur precisely because they are at this moment of change in a given variety (Moravcsik, 1978). This also invites us to include in the studies varieties that have been traditionally marginalized: creoles. As long as these types of

languages are incorporated in the different studies when they are, it is interesting to investigate them and see how they behave. The results may be comparable to the cases of borrowing that we find in the rest of the languages, and that can cause an exception to Universals, as Asia Minor Greek in Greenberg’s Universal 13 (Campbell et al., 1988) or other cases in Indo-European (Sapir, 1921). Croft (2003:232) understands that “languages do not occur in static or stable states. All languages exhibit some degree of grammatical variation, and they change over time”. This decision is something that, for example, is carried out by Mattiola (2020) in his study.

The previous view corresponds to a diachronic perspective of Universals, understood as invariances in the patterns of language change, the most common in Typology (Cristofaro, 2022). Within this, any claim based on current languages is assumed to be valid since it also applies to ancient languages. However, authors such as Bickel (2017a, 2017b) make explicit the excessive weight that certain areas may have for claims. What we believe to be Universal are simply preferences of a particular old area that have triumphed and spread to other areas and the rest of the world. This would be a possible explanation for the triumph of SOV or SVO over VOS or VSO. The first two orders have enjoyed greater success and dispersion throughout the world. According to this perspective, not interested in invariance in change, it is possible that we are losing information that did not succeed and disappeared. Grandi and Mauri (2022) also highlight the role of areas as fundamental, as they detail how in fragmented, small, and closed areas, languages with few speakers and less permeable are maintained. This causes greater richness in the linguistic type.

As will be seen in later sections and as commented by Guzmán Naranjo and Becker (2018) or Levshina (2019), the most settled perspective among researchers is processing preferences. In this sense, the previous explanation loses weight, and the weight that areas may have in this explication is reduced.

11.4 Scarcity of Data

Regardless of which of the three research methods is used to formulate Universals, the results may be highly influenced by the available data, which may be scarce and conditioned.

Most of the analyses of this problem in the bibliography focus on the bibliographic bias in the armchair methodology due to the logical prevalence of this type of study in Typology and because it has been the only existing method for many years. About questionnaires, it is also much easier to carry them out in places that coincide with a more significant presence of different grammars.

Likewise, the available corpora are scarce, and there is a tendency for the same languages to be represented as in previous approaches.

Rijkhoff and Bakker (1998) speak of bibliographic bias. The languages for which we have a better description are selected. Therefore, the scarcity of data is conditioning our selection and, thus, the knowledge of the world we have. Especially if we intend to extrapolate our claims from this selection as Universal.

This bibliographic bias is due to multiple factors. The lack of data for many languages lies in an imbalance in the strength of certain countries linked to some of these languages and the number of speakers. Odden (2003) states that 95% of the world's people speak 5% of the languages. This means, on the contrary, that 95% of the world's languages are spoken by 5% of the population. This data hampers the availability of data on this 95% of languages. Odden (2003), moreover, provides a list of many of the families that are undocumented for this reason, with many languages having fewer than 100,000 speakers. If there is documentation, in many cases, it is poor (word lists), which does not allow certain typological studies to be carried out on specific topics (Croft, 2003).

In short, it is the same position held by Cysouw (2009:2):

The presence of missing data is to a large extent inevitable in a typological survey. The available sources for a particular language are never complete, the more so when dealing primarily with strongly underdescribed languages, and it is not always possible (or feasible) to collect information directly through specialists or native speakers. The problem of missing data becomes more urgent when different features are combined into a single analysis. To be able to combine features, information on the same languages is needed for all features under investigation (though some corrections are possible to deal with missing data).

Therefore, if we do not have any information on more than two-thirds of the world's languages, the claims we make will always be about the same languages and extrapolating them to the rest (Bakker, 2010; Epps, 2010; Moravcsik, 2012).

Miestamo et al. (2016) go further and deal with problems of managing all this information. According to them, if we had good grammars for the 7000 languages of the world, it would not be possible to carry out these studies. First of all, he points out the difficulties of accessibility for researchers. The existence of these 7000 resources does not guarantee that a researcher can count on these data. If a researcher could do so, the study of elements in 7000 grammars, according to Miestamo et al. (2016), is too time-consuming and does not allow a complete study.

We would like to take advantage of this previous report to highlight the advantages of the new approach adopted in this thesis. Both data availability and efficient (non-time-consuming)

management are guaranteed in the case of corpus-based computational work. That is, if we had decent corpora in these 7000 languages in UD, for example, researchers all over the world would have equal access to the data (provided they have a free internet connection). The time spent would be almost the same, no matter if they are 70, 700 or 7000 languages. This is why we believe that more corpora should start to be produced in all languages with grammars, which would allow us to move forward more comfortably. However, Ponti et al. (2019) comment that the most extensive database (UD) is still showing a tiny size and should continue to grow.

It is also necessary to mention that the grammars approach is trying to mitigate all the existing access difficulties. Virk, Hammarström, Forsberg, and Wichmann (2020) talk about the intention to create a project (DReaM) that is a repository of the world's existing grammars. Some of the grammars are housed in private university libraries (this is the case for many Native American languages in the United States, for example). However, even now, access to the resource is still restricted, although it is hoped that it will soon allow free access to thousands of researchers who need it.

The next step currently being taken in this area of digitization of grammars is to enrich these texts to allow the use of enhanced search tools (Virk et al., 2020). These will aid systematic searches and allow the last of the problems pointed out by Miestamo et al. (2016) to be solved.

11.5 Anglocentrism

Another of the problems inherent to Linguistics is the risk we run when understanding a phenomenon influenced by the languages we know or which are of prestige. If some time ago this risk existed with Latin and Greek, even with French, nowadays English is the language with the most significant weight. In short, it is what some authors call WEIRD Languages (Henrich, Heine, & Norenzayan, 2010; Majid & Levinson, 2010): applying the sociological concept of Western Educated Industrialised Rich and Democratic Countries to Languages. There is an implicit assumption that the grammatical categories we use to understand these languages apply to any other language. In the case of Linguistic Typology, we must try to represent trends that apply to any language in the world. Therefore, it is essential that the concepts used are applicable and contain no bias. The awareness of this problem can already be seen in a clear way in Whorf (1956:241), who comments: “Our Indian languages show that with a suitable grammar, we may have intelligent sentences that cannot be broken into subjects and predicates. Any attempted breakup is a breakup of some

English translation or paraphrase of the sentence, not of the Indian sentence itself”.

We can also find this reflection in a debate originated by Greenberg himself. As Cooreman and Goyvaerts (1980) tells us, Greenberg claimed that Dutch was an SVO language, while Koster claimed that it was SOV. In reality, in these cases, what happens is that this language tolerates the classification of order in three constituents in a more complicated way because, in many cases, it duplicates and dislocates the verb. Applying this order, which works so well for a rigid language such as English, can be problematic for languages that function differently (Bybee, 2010; Whaley, 1997). However, in this case, we should note that Greenberg’s position seems more sensible if we stick to the requirements he puts forward for this basic word order, which more closely resembles the Dutch SVO (without fitting 100%, as it does in English). This aspect can be expanded in Siewierska (1988), Mithun (1992), and Song (2009).

Anglocentrism in the tools used to understand languages can extend to the world’s most widely used methods of analysis. For example, Comrie (1989:8) points out that X-theory is not good at understanding several languages in order to formulate Universals. This aspect has been contested by Coopmans (1983). However, it is easy to see the weight of Anglocentrism in this kind of analysis when the theories are primarily formulated from the English language. Occasionally, one or two other languages are added. However, this is always insufficient to capture a model that allows us to understand the behavior of any language in the world naturally and flexibly. This is something that is also clearly stated by (Daniel, 2010:3):

First, despite its Universalistic claim, in practice generative grammar has traditionally gravitated towards data from only a few of the world’s major languages. English provided the starting point for all generalizations. Once initial generalizations were produced, inclusion of non-English data led to slow modifications of the rules previously assumed to be Universal.

However, we can go further and talk not only about the models used as misleading but also about the terminology used and the concepts we use. We recommend checking out Bakker (2010) for more information on the influence of a generative vision when formulating claims or using certain conditioning labels. Dryer (1988a:197) clearly talks about the fact that:

A word is identified as an adjective if it expresses the kind of meaning associated with descriptive adjectives in English. It is clear, however, that in many languages, the words in question do not constitute a word class in the way they do in English. In many languages, the words in question are simply verbs; in other languages they are nouns. The use of the term adjective in word order studies is thus potentially misleading.

In the same sense, Gil (1997:188) exemplifies this by saying that “the imposition of familiar Eurocentric categories onto Riau Indonesian provides for an easy mode of grammatical description; however, it is essentially misguided”. If we highlight this quote, it is because it uncovers a

serious problem: grammatical description in these Anglocentric or Eurocentric terms is possible and may even be straightforward. However, it gives us a distorted picture of reality by forcing an interpretation through inappropriate lenses.

Problems of this kind have also been evident at other linguistic levels, such as pragmatics (P. Brown & Levinson, 1988; Hill, Ide, Ikuta, Kawasaki, & Ogino, 1986). For example, Wierzbicka (1985:145) points out that: “Features of English which have been claimed to be due to Universal principles of politeness are shown to be language-specific and culture-specific”. Fukada and Asato (2004) exemplify the inadequacy of the models of P. Brown and Levinson (1988) as Universal in application to specific languages such as Japanese. The most comprehensive critique of the confusion between Anglo-English and Universal pragmatics can be found in Wierzbicka (2006), which is the one we recommend for expansion. Therefore, as we indicated, it is a problem transversal to any cross-linguistic comparison.

Another quite obvious example is seen in the many criticisms of the implicational hierarchy of Berlin and Kay (1969). The criteria used to distinguish basic color terms work very well for languages like English; however, Saunders (1995) mentions brightness as the differential aspect in other languages and that it has been downplayed as a criterion. Roberson, Davidoff, and Davies (2000) add saturation as a differential aspect in some languages, such as Berinmo. This criticism is expanded in Levinson (2000).

It is, therefore, essential to allow a proper comparison between all the world’s languages by ensuring both a terminology applicable to all of them and a correct identification of the phenomena being searched for. However, it is difficult to detach ourselves from the categories we use. Dryer (1997) rightly defines them as a “convenient fiction”. Ramat (2009), for example, would not apply the concepts of “noun” or “verb”, as we know them, to Iroquoian languages. Adjectives, in the Eurocentric sense, also cause problems in some languages (Ramat, 2009).

This example of the adjective can be seen in the paper by Dryer (2019). This translates, at present, into the labeling as adjectives of elements with some verbal characteristics but which behave as adjectives.

Corbett (2009) shows how in the 21st century, there is already a desire to homogenize and propose standardizations within the conception of traditional labels in Linguistics to ensure a much less biased application in different languages. The most problematic aspect is how the different values that a language can have can be manifested. If we stay at a more general level and deal only with features, there are usually fewer problems (Corbett, 2009:137). Therefore, in order to ensure

comparability, all the more specific and characteristic aspects of a European conception of labels are discarded, and we are left with the more general and transversal aspects (Stassen, 2010).

In no case does the formation of new terms arise, but rather a reinterpretation of existing ones. This is due to what justifies (Bisang, 2010:1):

Terms such as ‘noun’ or ‘verb,’ which are found in almost any recent linguistic theory as well as in any descriptive grammar, are rooted in this tradition. Given the ubiquity of these terms, it is a fundamental task of Linguistics to analyze them and define them in a way that fits into our present-day knowledge about the range of cross-linguistic variation. This is the aim of a number of typologists.

This process, which began 20 years ago, is resolved by employing what is known as the comparative concept. We also find this reference as “*tertium comparationis*” in some authors, as it is understood as an element external to the languages analyzed, which allows us not to start from the reading of one of them that is applied to the other (Ponti et al., 2019). Or, in the words of Mauri (2022b:25-26):

un elemento terzo che resta costante e invariabile, rispetto al quale due o più lingue vengono confrontate. Il variare delle lingue viene in altre parole osservato rispetto a un concetto, un fenomeno, una funzione individuata in modo indipendente dalle lingue stesse, che costituisce il cuore dell’interrogativo di ricerca tipologico: come cambiano le lingue rispetto all’espressione del *tertium comparationis* individuato?

Haspelmath (2010:665) is concerned with defining them comprehensively:

Comparative concepts are concepts created by comparative linguists for the specific purpose of cross-linguistic comparison. Unlike descriptive categories, they are not part of particular language systems and are not needed by descriptive linguists or speakers. They are not psychologically real, and they cannot be right or wrong. They can only be more or less well suited to the task of permitting cross-linguistic comparison. They are often labeled in the same way as descriptive categories, but they stand in a many-to-many relationship with them [...]. Comparative concepts are Universally applicable, and they are defined on the basis of other Universally applicable concepts: Universal conceptual-semantic concepts, general formal concepts, and other comparative concepts.

In other words, the traditional descriptive categories are used, as we have already mentioned, but all the original characteristics that ascribed them to specific languages are eliminated (Masini & Mattioli, 2019). For example, in Becker (2021), the different elements in different languages that share domain, function, and systematic occurrence are understood as articles. This makes the comparison between languages possible and less forced. This does not mean that some difficulties do not remain and that all biases have been eliminated, which seems impossible for Bickel (2015). For him, these comparative concepts make it possible to analyze similar structures in the world’s languages but never identical ones.

Precisely, Mauri (2022a, 2022b) is a reference that does not treat “comparative concept” and “tertium comparationis” as synonyms precisely, contrary to most of the literature analyzed. In this case, tertium comparationis is understood as a concept with much more functional and semantically loaded characteristics: any syntactic issues are avoided. On the other hand, the “comparative concept”, for the cited reference, seeks a balance between semantic and morphosyntactic criteria that allow concepts to be defined not only by their function but also by a series of structural characteristics, as can be seen in the previous example of Becker (2021).

All these concerns have not been lost on those responsible for creating a Universal corpus annotation scheme for the Universal Dependencies repository: “The main goal is to provide a kind of lingua franca for grammatical annotation, which can be used for data interchange and development of multilingual systems” (Nivre, 2015:4). Therefore, the annotation guides consider all these aspects, allowing for a more natural application of linguistic concepts. Working with UD data ensures that this problem has been considered in the annotation guides and that the results are comparable.

11.6 Sampling

Studying specific characteristics in all the world languages (whether we refer to specific types or not) is impossible for various reasons. This forces us to select languages that represent the total number of languages in the world that we do not know. If we can verify the expected behavior of languages, the claims formulated will presumably apply to any equivalent language.

Traditionally, three different types of sampling are offered in Typology (Bakker, 2010; Guzmán Naranjo & Becker, 2022; Rijkhoff & Bakker, 1998):

1. Convenience Sampling

This occurs when the availability of data for a given phenomenon is inadequate and does not allow for Variety or Probability Sampling. A perfect balance in the sample cannot be guaranteed. However, the use of different types of languages (and the greater the convenience sampling, the greater the probability) allows results to be obtained that may be indicative of a Universal trend, which can be reinforced later by more precise sampling once more and better data are available.

Using a random selection of languages is discarded (Diessel & Hetterle, 2011; Rijkhoff & Bakker, 1998). This would not be genuinely random given the bibliographic bias we have

with the languages of the world; we would not be talking about representative randomness. Nor can we propose a selection of languages using traditional statistical criteria, as we cannot guarantee total independence of the languages without a prior review (Dryer, 1989). There may be genetic and area contamination, which Guzmán Naranjo and Becker (2022) name vertical and horizontal transmission, respectively.

2. Variety Sampling

It occurs when there is a wide variety of data from different languages, but the phenomenon has not been sufficiently studied in Typology. Suppose the information to be obtained is about the phenomenon itself and all its possible manifestations in the world, without considering its weight in Universal terms. In that case, this type of study is carried out. To this end, languages representative of the different linguistic types are selected. The representation of all families and genus and a good representation of areas are also considered. In addition to all this, a number of languages are often used with particular interest in this type of sample, as they are characteristically untypical (Miestamo et al., 2016). This means that balances or proportions between groups of languages are not respected, but instead, the aim is to capture the full spectrum of variation possible in a phenomenon. Hawkins (1980) proposes it as the best start to test a new approach.

3. Probability Sampling

Probability Sampling also occurs only when there is a wide variety of data from different languages. Commonly, this occurs when typological information on the phenomenon under study is already available but is to be weighted in terms of Universality. This check on the representativeness of a phenomenon can only be achieved through a balanced selection of languages. The balance must be respected in terms of language families, language areas, and the different linguistic types. The complexity of its design lies in the fact that the balance occurs in all senses simultaneously.

Despite the appropriateness of probability sampling for the study of Universals, it is not always possible to carry it out (Bakker, 2010). It must be understood that there is a high dependence on the object of study to be carried out. That is to say, although there may be data from a given set of languages that can be useful in sampling this type, we may not have the data that the researcher needs. Furthermore, Campbell et al. (1988) point out that it is impossible to maintain total neutrality or balance in samples and that samples are never perfect. We recommend extending this classification offered in Bell (1978).

In the case of the Universals formulated by Greenberg, they are based on data from 30 different languages. Of these, six are from the Indo-European family. On the other hand, there is no representation of Eskimo-Aleut, Caucasian, or Papua-New Guinean languages. This size cannot provide a representativeness that would allow for variety or probability sampling. The sample size is even smaller in the case of the formulation of implicational Universals. The languages analyzed are restricted to the only ones that meet the premise of the Universal. This, in the case of a Universal applicable to the VSO type, such as basic word order, can make the number very small, for example.

The languages were either known to the author, or the Universals were formulated on the basis of data taken from grammars. However, it must be said that there is a clear interest in balance and representativeness (although perfection cannot be guaranteed). In this selection of languages, languages from different areas of the world, languages of different morphological types, syntactic types, and families are added. A representative of each aspect is chosen (Greenberg, 1963).

Perkins (1989) proposes the range between 50 and 100 languages as sufficient to guarantee representativeness that offers acceptable claims in Typology. However, he also insists that it is advisable to subject the object of study to further research with other sample designs to confirm the claims made. Bakker (2010) also recommends that the study of Universals should be based on a sample of between 50 and 200 languages. A higher number of languages will not provide any substantially different or richer results and will instead compromise the independence of the selected languages in some cases. This is called by Guzmán Naranjo and Becker (2022) the oversampling problem. In any case, we believe that it makes more sense to carry out bottom-up sampling (sampling without thinking about the final number) than top-down sampling (setting the final number before starting) (Miestamo et al., 2016).

Dryer (1988a) puts his proposal into practice by analyzing 425 different languages, which will end up being 316 effectively. Dryer (1989) will go as far as 542 languages analyzed for certain aspects. However, the analysis of so many languages is not appreciated in the rest of the authors. The problem with such large samples, according to Rijkhoff and Bakker (1998), is that it is not possible to guarantee the absolute independence of so many languages from each other. The practical impossibility of achieving total independence in such extensive selections makes Piantadosi and Gibson (2014) argue that it is counterproductive and makes little sense to talk about or pretend to find absolute Universals. Moreover, such large samples fit better with variety sampling. In the case of probability sampling (for Universals), the information in the previous paragraph is understood as adequate.

In the late 1980s and 1990s, we found most of the innovative proposals on sampling. As mentioned above, the aim is to improve data availability from the grammar of the world's languages for the proposals of the different claims. Dryer (1988a) proposes two different ways of achieving balance from the genetic point of view (the most important at the time). One option is to collect one language for each of the existing families, 212 according to Dryer (2005). This language will be representative of its family. It will allow for balance in the case of high availability of languages of the same family, as all families will be equally represented (Merlo, 2015). This approach may not be entirely fair, depending on the language selected as representative of its family. The selected language may contain several typological features (for extralinguistic reasons, for example) that set it apart from the rest of the languages in that family.

Therefore, the second option, according to Dryer (1988a), is based on collecting multiple languages for each family. Once this data is available, the characteristics of these languages should be reviewed to decide what is most representative or characteristic of that family (or which language could be a good representative of this family). In this way, the balance between families will continue to be maintained. However, the effort and data availability required for this approach is greater. Moreover, as Comrie (1989) reminds us, one must not forget the balance between the representativeness of the universe and carrying out a study with manageable and practical data.

The balance of families can help us not to be conditioned by the size of certain families, which can cause us to have a distorted view according to Dryer (1989:259):

Imagine a world with 1000 languages. Suppose these 1000 languages are distributed over eleven language families such that there is one large family containing 900 languages and ten small families, each containing 10 languages. Suppose further that all 900 languages in the large family are SVO but that the languages in the ten small families are all SOV. I.e. in this world, 90% of the languages are SVO, while only 10% are SOV. What could we conclude about the possibility of a linguistic preference for SVO over SOV? Would the fact that 90% of the languages are SVO provide a reason for concluding that there is a linguistic preference for a SVO order? Clearly not.

In these years, we find a primacy for ensuring the representativeness of the different areas of the world. It is possible that the studies have many families, but that these do not respect the balance of the area, and there is an over-representation of certain areas of the world (Comrie, 1989).

For this reason, Dryer (1989) proposes the division of the world into five macro-areas, which he will later extend to six (Dryer, 1992): Africa, Eurasia, Papunesia, Australia, North America, and South America. The determination of these areas corresponds to a historical guarantee of no contact for as long as data has been available. This separation allows the claims to be as independent as possible. In the methodology proposed by Dryer (1989, 2005), checks are carried out within

the different areas. Thus, it is possible to check whether the exceptions to a Universal come from a specific area or whether, in reality, violations are distributed across different macro-areas. This also makes it possible to correct for the weight of areas such as Eurasia since whatever languages it has will have the same weight as Australia, for example, even if the number of documented languages is lower in percentage terms.

This proposal for macro-areas will remain quite limited to Dryer. We only found Hammarström and Donohue (2014) that try to work with it and present a slight modification in some of the borders drawn by Dryer to improve the independence of areas. There is also a proposal for areas in Bickel and Nichols (2013), which has not taken root and is not usually applied. Miestamo (2003, 2005) assumes the genus-macro-areas proposal as valid and will use it in his research without modification. In subsequent years, we will see other typologists who will continue to opt for sampling with the primacy of the genetic criterion, although with technical improvements in computation for selection (Rijkhoff et al., 1993). This is what is known as Diversity Value (Rijkhoff & Bakker, 1998): the selection of the languages within a genetic group that presents the greatest diversity among them. The researcher, therefore, no longer has to choose among all possible languages but among those that provide the highest Diversity Value (if their description in grammar is available).

Little will be said about the fact that within the typological study, other typological elements also have to be considered and controlled so that they cannot be misleading. Bakker (2010) clearly explains that if we want to formulate an implication of the type “If a language has Y, then it has Z” it is necessary to have a good balance and representation of Y, but also of X (the opposite type), to be able to determine the influence that Y has on Z, since perhaps in X it also behaves in this way and the restriction by means of the implication is worthless.

In addition, Bakker (2010) also presents a cultural bias that must be considered. He gives the example that the languages whose basic order has the object before the subject all have less than 3000 speakers. He believes that this may have some kind of unstudied relationship. However, this has not traditionally been taken into account by other typologists. We believe that the main problem behind this proposal is that the parameters and the real influences of this depending on the object of study are not easily objectifiable and, therefore, no corrective measure is applicable when creating a sampling beyond punctual considerations such as Bakker may make.

All these factors to be considered have been considered when creating samples for typological research. WALs has created a (variety) sampling of 200 languages to facilitate the study of researchers using this database. However, Hammarström (2009) points out that this creation has

no formal and rigorous set of procedures. In other words, the variety achieved with this sampling is questionable.

However, the proposal that seems most up-to-date and powerful (in the study through grammars) is that of Miestamo et al. (2016). It provides all the steps followed to achieve a detailed and rich sampling. One of its innovations is that it does not give importance to the language selected as representative of a genus, unlike in previous years. According to Miestamo et al. (2016), in samples as large as the ones we are working on, this possible adverse effect is diluted and balanced by other languages that already provide what, perhaps, might be missing. Mattiola (2020) is an example of research based on this proposal.

The last of the innovations in sampling is found in the proposal of Guzmán Naranjo and Becker (2022). This paper concludes that it is not necessary to use strict probability sampling to formulate Universals. Based on statistical tests they carry out, they show that a reasonably large sample covering the different areas and families is sufficient to propose valid claims in this sense. Therefore, once sufficient data are available on the object under study, substantially modifying the available languages is not necessary, as long as the minimum multi-area and multi-family representation mentioned above is guaranteed.

Although the criteria have been relaxed in recent years, it is not enough to consider the existing selection in UD as bias-free: “This is not a perfectly typologically balanced sample: Indo-European languages predominate” (Futrell et al., 2020:384). This database of different corpora still needs to grow to present some unrepresented families and areas. Once the under-represented or unrepresented parts are added, they will be balanced with the existing sufficient variety of Eurasia, and the claims will be free of bias. However, as seen in Jancso, Moran, and Stoll (2020), this sampling problem from the more computational perspective should not be understood as a call to give up research in this direction. If this were done, it would not be possible to continue to grow scientifically, and it would cost more to promote the changes indicated. Therefore, it is perhaps more convenient to carry out these studies with the understanding that all these problems continue to exist and the implications that this has.

UNIVERSITAT ROVIRA I VIRGILI

GLU - A GRADIENT MODEL FOR LANGUAGE UNIVERSALS. A COMPUTATIONAL REVISITATION OF GREENBERG UNIVERSALS

Antoni Brosa Rodríguez

Part III

Theoretical Framework

UNIVERSITAT ROVIRA I VIRGILI

GLU - A GRADIENT MODEL FOR LANGUAGE UNIVERSALS. A COMPUTATIONAL REVISITATION OF GREENBERG UNIVERSALS

Antoni Brosa Rodríguez

12

Typological Universals Frame

Contents

12.1	Data Type	159
12.2	Greenberg Universals	161

This thesis investigates Universals in language within the field of Linguistic Typology. Therefore, other possible perspectives on the concept of Universals in other disciplines or schools, such as those of Hockett (1963) or Chomsky (1965), are discarded. We want to arrive at a knowledge of the behavior of the world’s languages through the study of the manifestations of a wide variety of them (Newmeyer, 2008:35). In other words, a Language Universal is understood to be “a grammatical characteristic that can be reasonably hypothesized to be present in all or most human languages” (Moravcsik, 2010:1).

The fact that we are within Linguistic Typology implies a series of characteristics the study must have:

- **Comparison of the examples of multiple languages**

The study of Universals does not stem from introspection and general reflection on language. Nor is it based exclusively on the linguist’s knowledge of the languages he or she masters.

To be able to label the results obtained as plausible in the functioning of the world’s languages in general, they must arise from a large number of languages (Bakker, 2010). Moreover, they must not belong to the same area or have the same ancestry, as such parameters could influence the results.

- **Acceptance of exceptions**

It is to be expected and understood that there are languages that do not corroborate the theory supported by the vast majority of the world's languages. Results do not become invalid because they have counterexamples (Dryer, 1998).

This fact generates the appearance of labels such as statistical Universal, which reflect the validity of a Universal with exceptions. However, this is one of the points this study seeks to address, and a model to deal with it is proposed different from the prevailing one.

Aspects that could be improved in the discrete conception between absolute and statistical Universals are detected. There is also room for improvement in the distinction between a rule that can be useful for Linguistic Typology (Universal) and one that is not (non-Universal).

It has been seen in the different references in the bibliography that there is a certain discomfort with the theorization and classification of exceptions in Universals. Once we have seen that in Typology exceptions are accepted, we do not find any valid proposal that indicates in a natural way to what extent they should be accepted. In other words, what is the threshold between a statistical Universal and something not Universal. Moreover, we find it misleading that a Universal with 99.99% applicability shares a label with a Universal applicable to 73.44% of languages, for example. We consider the information of the 99.99% applicable Universal to be more comparable to that of a Universal with 100% applicability. This idea already appears in Jakobson (1963:212): “Uniformities with a probability slightly less than one are no less significant than uniformities with a probability of one”.

On the other hand, since we are working with hypotheses with representative samples of languages, it is pretty easy for an absolute Universal to become statistical in the future. We believe that paying attention and depending on whether an exception is Universal or not is not productive. However, it also seems unwise to set a hard threshold when 0.1% can change the conception of that Universal, mainly when that percentage depends on the study's sampling. This is why many typologists do not believe in the distinction between absolute and statistical (Dryer, 1998), although they continue to use these labels because there is no alternative.

In short, in this thesis, we not only accept the existence of exceptions to the formulated Universals as in any typological study, but we go further and see them as natural and gradual.

- **Increased knowledge of linguistic types**

Although the ultimate aim of Typology is the knowledge of the motivations of human language, the study of Universals focuses on describing the characteristics of different groups of languages (types). This explanation of the invariances within language variation is embodied in the formulation of rules that apply only to a set of languages that share some common grammatical feature (Grandi, 2020). This is why typological Universals are usually implicational.

- **Use of the comparative concept**

Since the research is carried out across multiple languages that are very disparate, the manifestation of grammatical elements may be nuanced or slightly different in any one of them (Haspelmath, 2010). Therefore, we work with comparative concepts, i.e., the labels used to analyze languages capture the essence of the object in question without going into precise or idiosyncratic aspects of that element in any given language. It is therefore necessary to have a joint parsing or annotation scheme for all languages proven to be general enough to be valid (Grandi & Mauri, 2022).

12.1 Data Type

Research in Linguistic Typology has traditionally been limited to comparing different language grammars. Many such studies have already been conducted to analyze more languages or discover new aspects of language behavior. Over the years, more languages have been described, and their description has become more and more effective for Linguistic Typology. It is easy to find the specific grammatical information that interests the linguist to compare it with the other languages (Levshina, 2022).

Although this method is the one that allows the best sampling of languages, this type of approach has several limitations:

- Some linguists may find it difficult to have all the grammars available.
- There is no homogenization in terminology and analysis of grammars.
- There may be linguistic elements that are not represented or commented on.
- Borderline or concrete structures may not be represented.

- Information is reduced to a natural language label.

We have also seen how obtaining linguistic information through questionnaires is possible. In this case, obtaining the information of interest to the linguist alone would be possible. There is also a homogenization of the method of collection and analysis. However, there is no guarantee of the neutrality of the data obtained, which have been elicited. The determining factor that hinders using questionnaires for linguistic analysis is a practical one. On the one hand, if one wants to carry out an in-depth study allowing the collection of data relating to 45 different Universals, for example, a large amount of information has to be obtained. The size of such a questionnaire may not be practical if several occurrences and different levels of detail are to be obtained. On the other hand, it is too logistically and economically demanding. The linguist in charge must ensure the questionnaire is implemented on a representative sample of subjects, not from one language but tens or hundreds of languages. Collecting such data in remote, inaccessible, and other known constraints is very difficult.

For this reason, the study of linguistics through computation has been gaining importance in recent years. This methodology will be used in this thesis, and the data will come from corpora. In this case, we must recognize a limitation about the number of languages with corpora available. In the tool we will use, there are about 150. We believe this allows an adequate study in Typology, although it cannot be compared with the variety in grammar studies. However, we believe that this more limited availability of data can be reversed simply with the passage of time and a more significant commitment to it. In return, we see some benefits in its adoption:

- Data are available on the Internet in open access.
- There are standardized annotation schemes. The linguist does the analysis.
- The linguist is the one who looks for the specific linguistic elements in the language texts.
- Borderline structures, or structures that are less frequent but exist in a language can be captured.
- Quantitative results can be obtained that show a greater gradation within the basic linguistic types.

In order to be able to process the data obtained and convert it into meaningful information for the Typology and to favor the comparison with the classic labels, the proposals of Levshina (2019, 2022); Levshina et al. (2021), and Gries (2013) are followed.

12.2 Greenberg Universals

This thesis is intended to be a methodological review, both theoretical and practical. The theoretical part refers to the classification of Universals. The practical part corresponds to revisiting the 45 Universals of Greenberg (1963) with another perspective through a computational and quantitative approach. There are several reasons for this choice:

1. It is the most representative and famous proposal since it is the origin of this discipline and Universals as we understand them today. This also implies that the other proposals, such as those of Dryer (1992, 1997, 2013); Lehmann (1978); Vennemann (1974, 1976) are based on the fundamental guidelines of Greenberg’s work. They are a particular update of very few specific Universals.
2. We understand the proposal to be the most challenging. The rest of the proposals are shorter and more limited, focused on a specific and general aspect of languages—for example, the correspondence between the direction of different syntactic elements (word order). In contrast, Greenberg offers a complete overview of different morphosyntactic features, both general and concrete, based not only on word order. This implies a greater difficulty of formalization since, in some Universals, the pattern described is truly complex.
3. His proposal allows us to work with up to 45 different Universals, while others are simply a single Universal in a hierarchy format. Therefore, we believe it is the best theoretical framework to put our knowledge into practice and that, after this doctoral thesis, we will be able to apply it to any proposal by any author.

It is also necessary to mention that, to our knowledge, there has not yet been any attempt to test and reformulate Greenberg’s Universals from the use of language using computational and corpus linguistics in the framework of Universal Dependencies. The only thing we have found, as we mentioned, is some attempt (in the last five years, a fact that makes us suspect that it is a topic of interest for the scientific community) to analyze some specific Greenberg Universal, primarily related to the basic order of the subject, verb and object of languages (Brosa-Rodríguez & Jiménez-López, 2023; Choi, Guillaume, & Fort, 2021; Choi, Guillaume, Fort, & Perrier, 2021; Gerdes, Kahane, & Chen, 2019; Yan & Liu, 2023). In no case we have seen either an analysis of the totality of Greenberg’s Universals, or a proposal for reformulation or extension of these Universals, or even a reflection on which Universals are technically verifiable and which are not.

Moreover, in the case of the already analyzed Universals that we have mentioned, we must say that we believe that they have not been formalized correctly for the reasons we have already explained and, therefore, we propose different patterns, which allow us to have an alternative and (in our opinion, more accurate) approach to Greenberg's Universals.

These 45 Universals we work with correspond to surface or typological Universals. Most (40) are implicational, while very few are unrestricted (5). Regarding the linguistic level of the 45 Universals, 28 involve some aspect of word order, 16 are exclusively about morphology or features, and 1 is about intonational patterns. The original study of Greenberg (1963) is based on 30 languages: English, Basque, Serbian, Welsh, Norwegian, Modern Greek, Italian, Finnish, Yoruba, Nubian, Swahili, Fulani, Masai, Songhai, Berber, Turkish, Hebrew, Burushaski, Hindi, Kannada, Japanese, Thai, Burmese, Malay, Maori, Loritja, Maya Zapotec, Quechua, Chibcha, Guarani. These languages are either known to Greenberg, or he has relied on the information in grammars. Thus, there is a mix between the convenience of the availability of the sources and an effort to reflect languages of very different origins, areas, and characteristics.

These 45 Universals are:

Universal 1: In declarative sentences with nominal subject and object, the dominant order is almost always one in which the subject precedes the object.

Universal 2: In languages with prepositions, the genitive almost always follows the governing noun, while in languages with postpositions it almost always precedes.

Universal 3: Languages with dominant VSO order are always prepositional.

Universal 4: With overwhelmingly greater than chance frequency, languages with normal SOV order are postpositional.

Universal 5: If a language has dominant SOV order and the genitive follows the governing noun, then the adjective likewise follows the noun.

Universal 6: All languages with dominant VSO order have SVO as an alternative or as the only alternative basic order.

Universal 7: If in a language with dominant SOV order, there is no alternative basic order, or only OSV as the alternative, then all adverbial modifiers of the verb likewise precede the verb.

Universal 8: When a yes-no question is differentiated from the corresponding assertion by an intonational pattern, the distinctive intonational features of each of these patterns are reckoned

from the end of the sentence rather than from the beginning.

Universal 9: With well more than chance frequency, when question particles or affixes are specified in position by reference to the sentence as a whole, if initial, such elements are found in prepositional languages, and, if final, in postpositional.

Universal 10: Question particles or affixes, when specified in position by reference to a particular word in the sentence, almost always follow that word. Such particles do not occur in languages with dominant order VSO.

Universal 11: Inversion of statement order so that verb precedes subject only occurs in languages where the question word or phrase is normally initial. This same inversion occurs in yes-no questions only if it also occurs in interrogative word questions.

Universal 12: If a language has dominant order VSO in declarative sentences, it always puts interrogative words or phrases first in interrogative word questions; if it has dominant order SOV in declarative sentences, there is never such an invariant rule.

Universal 13: If the nominal object always precedes the verb, then verb forms subordinate to the main verb also precede it.

Universal 14: In conditional statements, the conditional clause precedes the conclusion as the normal order in all languages.

Universal 15: In expressions of volition and purpose, a subordinate verbal form always follows the main verb as the normal order except in those languages in which the nominal object always precedes the verb.

Universal 16: In languages with dominant order VSO, an inflected auxiliary always precedes the main verb. In languages with dominant order SOV, an inflected auxiliary always follows the main verb.

Universal 17: With overwhelmingly more than chance frequency, languages with dominant order VSO have the adjective after the noun.

Universal 18: When the descriptive adjective precedes the noun, the demonstrative, and the numeral, with overwhelmingly more than chance frequency, does likewise.

Universal 19: When the general rule is that the descriptive adjective follows, there may be a minority of adjectives which usually precede, but when the general rule is that descriptive

adjectives precede, there are no exceptions.

Universal 20: When any or all of the items (demonstrative, numeral, and descriptive adjective) precede the noun, they are always found in that order. If they follow, the order is either the same or its exact opposite.

Universal 21: If some or all adverbs follow the adjective they modify, then the language is one in which the qualifying adjective follows the noun and verb precedes its nominal object as the dominant order.

Universal 22: If in comparisons of superiority, the only order, or one of the alternative orders, is standard-marker-adjective, then the language is postpositional. With overwhelmingly more than chance frequency if the only order is adjective-marker-standard, the language is prepositional.

Universal 23: If in apposition the proper noun usually precedes the common noun, then the language is one in which the governing noun precedes its dependent genitive. With much better than chance frequency, if the common noun usually precedes the proper noun, the dependent genitive precedes its governing noun.

Universal 24: If the relative expression precedes the noun either as the only construction or as an alternate construction, either the language is postpositional, or the adjective precedes the noun or both.

Universal 25: If the pronominal object follows the verb, so does the nominal object.

Universal 26: If a language has discontinuous affixes, it always has either prefixing or suffixing or both.

Universal 27: If a language is exclusively suffixing, it is postpositional; if it is exclusively prefixing, it is prepositional.

Universal 28: If both the derivation and inflection follow the root, or they both precede the root, the derivation is always between the root and the inflection.

Universal 29: If a language has inflection, it always has derivation.

Universal 30: If the verb has categories of person-number or if it has categories of gender, it always has tense-mode categories.

Universal 31: If either the subject or object noun agrees with the verb in gender, then the adjective always agrees with the noun in gender.

Universal 32: Whenever the verb agrees with a nominal subject or nominal object in gender, it also agrees in number.

Universal 33: When number agreement between the noun and verb is suspended and the rules is based on order, the case is always one in which the verb precedes and the verb is in the singular.

Universal 34: No language has a trial number unless it has a dual. No language has a dual unless it has a plural.

Universal 35: There is no language in which the plural does not have some non-zero allomorphs, whereas there are languages in which the singular is expressed only by zero. The dual and the trial are almost never expressed only by zero.

Universal 36: If a language has the category of gender, it always has the category of number.

Universal 37: A language never has more gender categories in non-singular numbers than in the singular.

Universal 38: Where there is a case system, the only case which ever has only zero allomorphs is the one which includes among its meanings that of the subject of the intransitive verb.

Universal 39: Where morphemes of both number and case are present and both follow or both precede the noun base, the expression of number almost always comes between the noun base and the expression of case.

Universal 40: When the adjective follows the noun, the adjective expresses all the inflectional categories of the noun. In such cases the noun may lack over expression of one or all of these categories.

Universal 41: If in a language the verb follows both the nominal subject and nominal object as the dominant order, the language almost always has a case system.

Universal 42: All languages have pronominal categories involving at least three persons and two numbers.

Universal 43: If a language has gender categories in the noun, it has gender categories in the pronoun.

Universal 44: If a language has gender distinctions in the first person, it always has gender distinctions in the second or third person, or in both.

Universal 45: If there are any gender distinctions in the plural of the pronoun, there are some gender distinctions in the singular also.

13

Token-Based Typology

Contents

13.1	Universal Dependencies Project	168
13.2	Benefits of Universal Dependencies	171
13.3	Limitations of Universal Dependencies	172
13.4	Tools to Analyze Data from Universal Dependencies	174

In Typology, computational approaches are often linked to large databases, i.e., corpora. Thanks to technological advances, it is possible to consult precise information in millions of data efficiently and quickly. In the case of the interest in formulating Universals, it is necessary to consider that it is necessary to work with different corpora from different languages. Therefore, there is a need for comparison, which means that the corpora allow it.

These corpora can be, fundamentally, of the following types (Levshina, 2022):

1. Parallel corpora

This type of text has a greater representation of languages. The most translated and available text is the New Testament. The sentences of the translations are aligned with English, which allows for comparison. However, the relevant linguistic information is not labeled. Therefore, even if the translation is available, sentence by sentence, the information in unknown languages is useless for a typological study.

Recently, more projects such as PartTy (Parallel Corpus for Typology) have appeared. In this case, we are dealing with aligned translations of different movies. Again, we do not have precise grammatical information. In addition, the process is automatic, which may mean some errors.

Some parallel-aligned corpora have more (but not excessive) grammatical information, such as the European Union Acts. However, it is not appropriate to use them because of the bias of the area used, which is only one continent.

2. Comparable corpora

In this case, we are at an intermediate point between parallel corpora and corpora with unified annotation. These are corpora of different languages that are based on the same subject. A very clear example of this is Multi-CAST. In this case, there is an annotation of grammatical elements. However, only a dozen languages are available, which makes typological studies difficult. New projects such as DoReCo, which focuses on phonology and has 50 languages, are under development or have just been published. Other linguistic information is not as detailed as in other proposals. In this case, it is based on oral narratives.

3. Corpora with unified annotation

There are corpora of different languages and subjects that share the same grammatical annotation scheme. This variety allows for claims that are much less restricted to textual modality. It also allows for greater flexibility in incorporating different languages. The annotation scheme ensures comparability between the different language corpora and facilitates their reading and processing. The most representative example is Universal Dependencies, a tool created for such purposes.

Therefore, different sources of data for language comparisons are constantly evolving. The choice of one or another depends on the type of study to be carried out. However, the tool that has evolved the most and currently provides the most data to carry out a relatively reliable study is Universal Dependencies.

13.1 Universal Dependencies Project

Universal Dependencies (UD) is an open repository of corpora (Nivre et al., 2023). The interest in this repository compared to other options is due to its broad representation of languages. The most recent version (2.11) has 243 treebanks, corresponding to almost 150 different languages. Although WEIRD (Western Educated Industrialised Rich and Democratic) languages are more widely represented and have more extensive and more varied corpora, there is a genuine interest in offering the same resource with less successful languages from all over the world. The tool

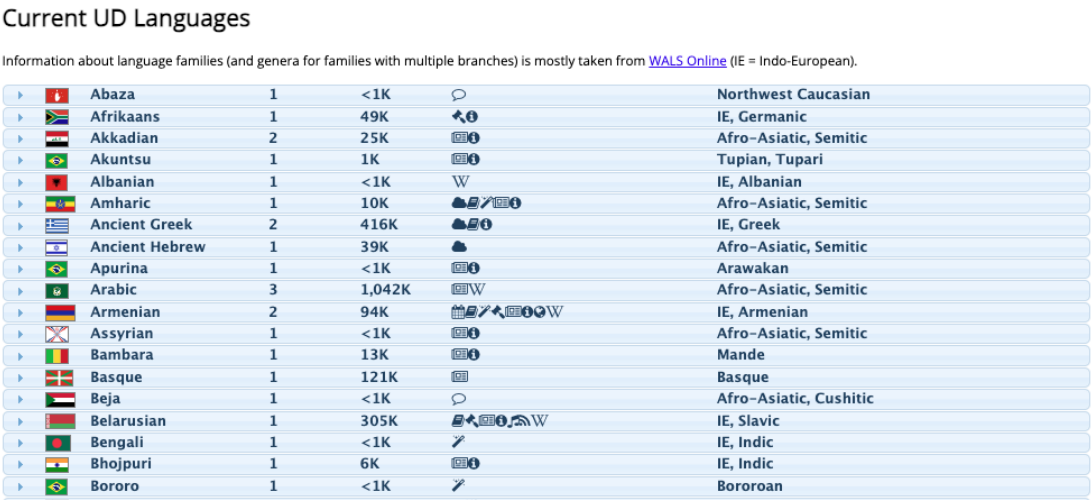


Figure 13.1: Universal Dependencies Website

constantly expands to include more treebanks and languages, allowing for an increasingly exciting representation of the world’s languages. An example of this can be seen in the figure 13.1.

The interest of the tool lies fundamentally in the two words that form its title. Firstly, it claims to be Universal. In other words, it aims to accommodate (potentially) any language worldwide. To this end, it must guarantee a series of particular annotation criteria, which we will see later. This ensures comparability between the corpora of different languages. Secondly, the syntactic annotation scheme chosen is dependencies. This is due to the realization that an annotation using the framework of Tesnière (1959) is much more computationally fruitful: it is more straightforward. Given the simplicity of the parsing steps, it offers a higher level of correctness.

UD’s managers point out that (Nivre et al., 2023):

1. UD needs to be satisfactory for analysis of individual languages.
2. UD needs to be good for Linguistic Typology.
3. UD must be suitable for rapid, consistent annotation.
4. UD must be suitable for computer parsing with high accuracy.
5. UD must be easily comprehended and used by a non-linguist.
6. UD must provide good support for downstream NLP tasks.

The problem with these objectives is that, although they are complementary, in some cases, they may require conflicting approaches. Therefore, we see it as a significant challenge to strike a

```
# sent_id = ParlaMint-IT_2020-03-11-LEG18-Sed-200_192
# text = Il diritto alla dignità e alla vita dei detenuti vale tanto quello di tutti gli altri cittadini, e lei lo sa benissimo.
1  Il      il      DET      RD      Definite=Def|Gender=Masc|Number=Sing|PronType=Art      2      det      -      -
2  diritto diritto NOUN      S      Gender=Masc|Number=Sing      13      nsubj      -      -
3-4  alla    -      -      -      -      5      case      -      -
3  a      a      -      -      -      -      -      -
4  la      il      DET      RD      Definite=Def|Gender=Fem|Number=Sing|PronType=Art      5      det      -      -
5  dignità dignità NOUN      S      Gender=Fem      2      nmod      -      -
6  e      e      CCONJ    CC      -      9      cc      -      -
7-8  alla    -      -      -      -      -      -      -
7  a      a      -      -      -      -      -      -
8  la      il      DET      RD      Definite=Def|Gender=Fem|Number=Sing|PronType=Art      9      det      -      -
9  vita    vita    NOUN      S      Gender=Fem|Number=Sing      5      conj      -      -
10-11 dei      -      -      -      -      -      -      -
10 di      di      -      -      -      -      -      -
11 i      il      DET      RD      Definite=Def|Gender=Masc|Number=Plur|PronType=Art      12      det      -      -
12 detenuti detenuto NOUN      S      Gender=Masc|Number=Plur      2      nmod      -      -
13 vale    valere  VERB      V      Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin      0      root      -      -
14 tanto  tanto  ADV      B      -      13      advmod      -      -
15 quello quello  PRON      PD      -      13      obj      -      -
16 di      di      -      -      -      -      -      -
17 tutti  tutto  DET      T      -      20      case      -      -
18 gli    il      DET      RD      Definite=Def|Gender=Masc|Number=Plur|PronType=Art      20      det      -      -
19 altri  altro  ADJ      A      Gender=Masc|Number=Plur      20      amod      -      -
20 cittadini cittadino NOUN      S      Gender=Masc|Number=Plur      15      nmod      -      SpaceAfter=No
21 ,      ,      PUNCT    FF      -      25      punct      -      -
22 e      e      CCONJ    CC      -      25      cc      -      -
23 lei    lei     PRON      PE      Numbers=Sing|Person=3|PronType=Prs      25      nsubj      -      -
24 lo     lo      PRON      PC      Clitic=Yes|Gender=Masc|Number=Sing|Person=3|PronType=Prs      25      obj      -      -
25 sa     sapere  VERB      V      Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin      13      conj      -      -
26 benissimo bene    ADV      -      Degree=Abs      25      advmod      -      SpaceAfter=No
27 .      .      PUNCT    FS      -      13      punct      -      SpacesAfter=\n
```

Figure 13.2: Treebank Annotation Example in Raw

balance to make all these tasks possible.

UD corpora can be annotated directly in UD format or corpora with a previous annotation that has been converted. In any case, the information contained (in a unified way) is always the same, as exemplified in figure 13.2.

In this organization, by columns, we have the information related to:

1. the form of the word
2. its lemma
3. its Part of speech in UD format
4. its Part of speech in original annotation (if exists)
5. its morphological features
6. the number corresponding to the word on which the word depends
7. its syntactic function

By annotating all this data, different queries can be carried out to extract the invariants in the element of interest to the researcher.

13.2 Benefits of Universal Dependencies

Most of the advantages of this tool are shown at different points in this thesis and are the following:

1. **Having millions of data annotated with quickly processed dependencies allows for a different approach than the traditional Typology.**

We can work efficiently with labeled corpora. In Typology, the approach to Universals in language prefers data analysis to other methodologies, such as abstraction. We believe it is ideal for the methodology in Typology to use corpora for this purpose. If the available data is expanded, the script with the formalizations made can be reused and applied to the new corpora.

2. **We do not have second-hand information simplified in a label (which is often unrepresentative).**

In this case, we have information based on real text data that allows for greater granularity. It may be that the analysis of some language in a grammar is erroneous in some aspects, and such an error is perpetuated. It may be that the analysis of the grammar of a language has been influenced by a particular linguistic view that may make comparison with other languages difficult. In any case, moreover, grammars contain only language information under a label. In the case of UD, it allows the most precise knowledge of the behavior of such a language. An SVO language like Russian is not equivalent to an SVO language like English. The syntactic rigidity of the latter makes it not comparable to the former, although both are collected under the SVO type.

3. **This level of detail makes it possible to arrive at more precise claims and bring linguistic phenomena closer to the continuum they represent.**

The possibility of seeing the concrete occurrences of a phenomenon in each language subsequently allows grammatical phenomena to be viewed as a continuum. This favors claims such as that of Levshina et al. (2021).

4. **This arrangement of data also makes it possible to capture borderline or non-prototypical cases in the given language.**

Grammars usually include only the general behavior of the language under analysis. Information on non-prototypical elements that exist in the language does not usually appear.

For example, in a grammar we can find the information that a given language is prepositional. On the other hand, in UD, we can obtain the total occurrences of prepositions and postpositions existing in a language. This makes it possible, for example, to detect postpositions in English, such as “ago”, which might otherwise go unnoticed in less studied languages.

5. **The quantitative approach offers the possibility of studies based on new, more precise, or different metrics, as in the case already shown of Futrell et al. (2020, 2015).**

It is understood that a quantitative approach offers fewer doubts than an approach carried out with a greater weight of introspection or the use of grammars, which can generate some kind of bias. It is also difficult to compare data from different languages if they have been translated into grammars using different labels. The lack of unification can make comparability difficult or misleading. This is also a risk when using data for a specific language from an author strongly influenced by any linguistic perspective.

6. **Free and immediate availability of all UD data.**

We have compared this project with approaches to the same phenomenon through laboratory tests or by consulting grammars, and it only has advantages from a technical (ease and effectiveness of data processing) and ethical (open science and delocalization of the availability of resources) point of view.

13.3 Limitations of Universal Dependencies

However, a number of limitations still exist in UD that need to be mentioned to warn the reader of possible biases. These limitations are the following:

1. **The annotation of the different corpora by different people.**

This means that there may be disparities in some of the criteria used. However, two fundamental aspects should be emphasized. This problem also exists in the Universals formulated by grammars, as no linguist is an expert in all the languages analyzed. Moreover, the annotation guides that have been developed, with a high level of detail, are designed to mitigate this aspect as much as possible and to ensure that the same criteria are applied. Therefore, we believe that this is the best alternative within the possibilities.

2. The detailed annotation scheme does not exempt its implementation from being perfect.

We have found some examples of specific errors in some treebanks. These are detailed in the corresponding Universal and can be summarised in two broad types. We use the origin of the error to classify them. On the one hand, we have detected some cases of lack of information. Even though a language has a morphological feature, and this is in the annotation guide, some annotators (if it is manual) may forget to add it, which would generate the false conception that there is no such value. On the other hand, it is also possible to find the opposite: an overgeneration of non-factual information. This appears especially when the corpus has been annotated with another scheme and converted to UD. In that case, the process may add features to words that do not have them because they coincide with another part of speech that does exhibit them. If serious supervision is not carried out after conversion, we may be pointing out things that do not exist. Nevertheless, the errors are particular to particular corpora. We believe that the benefits outweigh these one-off errors (which are easily rectified). Therefore, we believe that the most sensible thing to do is to point out these errors but not to give up investigating the phenomena, as biases can occur in any approach; there is no perfect method.

3. Sampling is limited by the source used.

The available languages are much more limited compared to the grammars. An essential part of these are from the Indo-European family, and there is a clear tendency to overrepresent Eurasian languages compared to other areas, such as Australia. Adequate sampling is, therefore, not possible. This is why we prefer to analyze all available languages. This will at least allow us to obtain a typological representation of as many languages as possible, even if many belong to the same language family. We believe that it is essential not to forget this aspect since many of these languages have not been analyzed from the perspective raised here in some aspects. In the future, with a broader representation and with these languages already analyzed, perhaps some specific selection would be more interesting.

4. The same bias can be applied to the textual Typology of the treebanks available for some languages.

This fact is seen in dead languages, usually religious texts. However, in this case, we must point out that these types of texts are also used in grammar, as there is usually no other type of witness for that language. It is quite true that the typological characterization of a language may be influenced by the sole availability of literary texts of that language, which

may be faithful reflections of it. However, the alternative would be not to carry out this research. Therefore, we believe it is more interesting to propose claims based on these texts rather than give up investigating this fact. The influence of the type of text is real but limited; for some specific aspects that are analyzed (for example, whether a language is gendered or not), we do not believe that the influence is relevant.

5. **The convenience sampling and textual typology biases must be mitigated.**

On the one hand, we have merged statistical Universals with absolute Universals. On the other hand, we renounce a typological characterization based on exact metrics that we believe may be misleading insofar as they are based on data that may be biased (as recent studies have attempted to do). A more general characterization, using a model based on Fuzzy Logic to determine whether a claim is interesting for the Typology or not (High-Universal), serves to reduce the impact of a slight error: the result in the case of slightly modifying the number of exceptions is the same.

6. **The data provided in UD is raw.**

The repository only hosts files with the appearance shown in previous pages in the figure 13.2. In order to compare the results between different files (different treebanks and languages) and extract the exact information of the desired syntactic structures, we cannot do this manually by working with millions of data. This data needs to be processed in a computational way, which requires the linguist to acquire or apply such knowledge to the tools designed for it. Since UD does not have this, it recommends different tools created for this purpose, which are analyzed in the following section, allowing us to assess what is most suitable for our study.

13.4 Tools to Analyze Data from Universal Dependencies

To work with the data present in Universal Dependencies (De Marneffe et al., 2021), we have five different tools endorsed by the framework itself (Nivre et al., 2023). These are:

1. Scalable and Efficient Tree Search (SETS) (Luotolahti, Kanerva, Pyysalo, & Ginter, 2015)
2. Prague Markup Language - Tree Query (PML Tree Query) (Štěpánek & Pajas, 2010)
3. Kontext (Machálek, 2020)

4. Graph Rewriting for NLP - Match (Grew-Match) (Guillaume, 2021; Perrier & Guillaume, 2019)
5. Infrastructure for the Exploration of Syntax and Semantics (INESS) (Rosén, De Smedt, Meurer, & Dyvik, 2012; Rosén, Dyvik, Meurer, & De Smedt, 2017)

Each has distinct positive and negative points. Therefore, it is worth analyzing all of them to see which is most suitable for the research we want to conduct in this study. After defining and briefly explaining the functioning of each of the tools, we will highlight precisely their positive and negative points. In the case of the first tool we have analyzed, this is not possible, and we have included and explained this in the “inoperability” section.

13.4.1 Scalable and Efficient Tree Search

This tool was created by researchers at the University of Turku. Its web interface allows the researcher to perform the queries online faster. It is also important to note that it is open-licensed (but, actually, all the presented tools are).

Inoperability

Once you enter the website, you can quickly see the shortcomings of this tool. First, the page where the “query language” is explained, i.e., the formal language used to perform uniform searches, is broken and does not allow access. This fact already makes using Scalable and Efficient Tree Search (SETS) impossible since the researcher does not know the full search potential of the tool (what it allows, what it does not allow, and how it is done).

On the date of the query (February 9, 2023), it can be seen that the tool does not work. The default query (`<nsubj_`) is performed in the available treebanks and constantly loads without showing any results.

It should also be noted that it works only with the raw data from Universal Dependencies. It does not work with other parallel annotation models that allow comparisons, such as Surface Universal Dependencies (SUD) or Enhanced Universal Dependencies, which are proposals aimed at providing greater detail in the annotation. A summary of the simplicity of the resource can be seen in figure 13.3.

For all these reasons, the assessment of the use of SETS for research cannot be continued.

Query

Result Limit:

☐ Case Insensitive

Query Language is documented [here](#)

Databases

- PB
- ▾ ☒ UD
 - ▾ ☒ UD_Afrikaans-AfriBooms
 - ▾ ☒ UD_Akkadian-PISANDUB
 - ▾ ☒ UD_Albanian-TSA
 - ▾ ☒ UD_Amharic-ATT
 - ▾ ☒ UD_Ancient_Greek-PROIEL
 - ▾ ☒ UD_Ancient_Greek-Perseus
 - ▾ ☒ UD_Arabic-NYUAD
 - ▾ ☒ UD_Arabic-PADT
 - ▾ ☒ UD_Arabic-PUD
 - ▾ ☒ UD_Armenian-ArmTDP
 - ▾ ☒ UD_Assyrian-AS
 - ▾ ☒ UD_Bambara-CRB
 - ▾ ☒ UD_Basque-BDT
 - ▾ ☒ UD_Belarusian-HSE
 - ▾ ☒ UD_Bhojpuri-BHTB
 - ▾ ☒ UD_Breton-KEB
 - ▾ ☒ UD_Bulgarian-BTB

Figure 13.3: General View of “SETS Treebank Search” Tool

13.4.2 Prague Markup Language - Tree Query

This tool is hosted at Lindat Clarin, which means that it is promoted by the UFAL (Institute of Formal and Applied Linguistics and the Computer Science School, Faculty of Mathematics and Physics, Charles University, Czech Republic). That is, from one of the hearts of Universal Dependencies. It is a tool that has been active for over a decade; although this does not mean it is obsolete, it looks wholly updated and functional. There is a web version and a server download version. However, the web version only contains some of the full functionality. The interface design is very clear, clean, and intuitive. Another positive aspect to highlight is the large number of hosted corpora. These are the corpora that can be found in Universal Dependencies. However, they have decided to incorporate all the existing versions of the same corpus separately, which facilitates specific searches in previous versions (if the researcher needs it). This can be seen in figure 13.4.

Positive points

The way of writing the queries is relatively formalized, which allows a clear language, replicable in any language, and practical. The selection of the various elements of the query can be done either by selecting the syntax to be used from the top drop-down menu or by typing directly in the text box. One of the clear conveniences, especially for a linguist unaccustomed to the computational

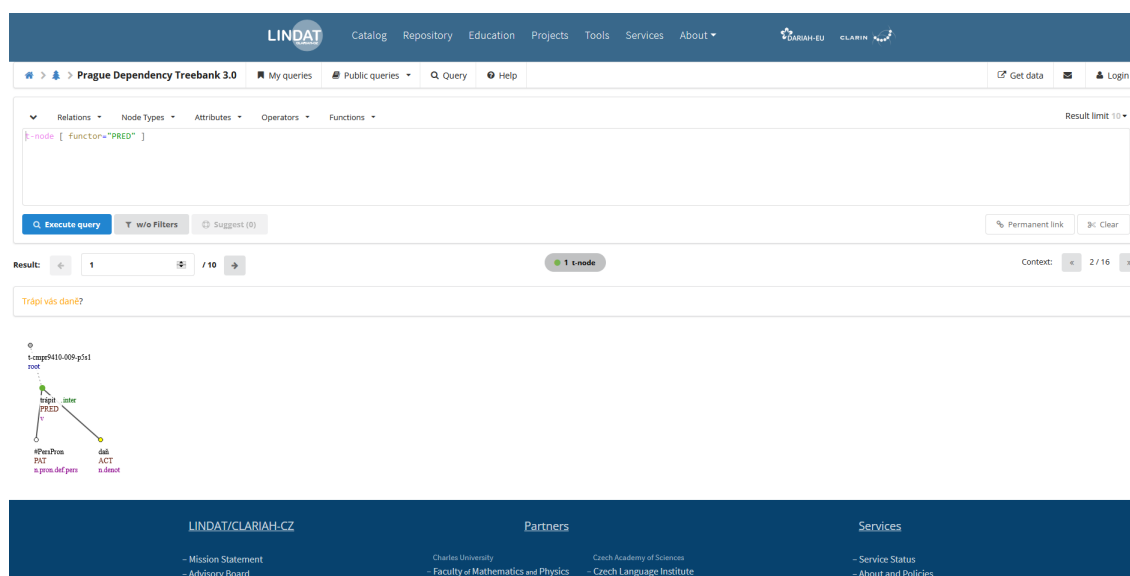


Figure 13.4: General View of “PML Tree Query” Tool

use of corpora, is that the various words in the query are auto-completed as you start typing, allowing you to avoid various typing errors in the formalism. However, the nomenclature and the large number of different menus can be confusing for the linguist. Another positive aspect is that it allows the search for any data annotated in such corpora, as well as for aspects that are more in the semantic domain (such as the role of the agent). This information is annotated as an extra in some corpora.

Finally, we would also like to highlight that it has a tutorial, allowing us to see the different possible queries. However, none of the syntax that can be used between the different treebanks and the combinatorial potential of these for different constraints appear systematically or clearly. Some isolated information appears within each treebank of each language, but it is neither exhaustive nor Universal and applicable to all languages. Therefore, we believe it is a tool designed to analyze specific cases of specific languages and not so much for typological or Universal studies.

Negative points

As for the negative aspects, we must also highlight several things. First, most interesting semantic queries are only valid for English and Czech (the favored languages in UFAL and all its tools). It is practically impossible to obtain results in significant, essential languages such as Spanish and Russian. Moreover, as seen from its tutorial, it is somewhat complex to follow; it is a tool more focused on extracting as much semantic information as possible. As we have commented above, this only works in a solvent way for Czech and English, so no other language is available

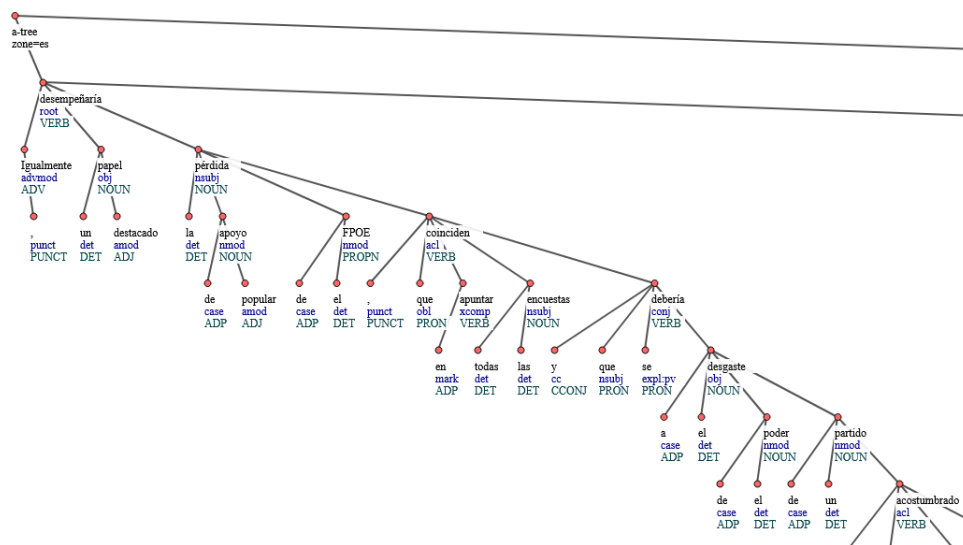


Figure 13.5: Query in “PML Tree Query” Tool

(in short, no typological or Universal studies can be made of this). We would also like to take this opportunity to mention that the guide itself is focused on explaining in a detailed way the queries related to semantics (for English and Czech), forgetting the other languages and the different possibilities offered.

As for the data representation (remember that the web version does not allow graphs), it is pretty complicated to read. This is because they have opted for a representation of the dependencies in tree format and not in arrow format, which makes it difficult to read as it is pervasive. For example, we offer the first sentence we obtain when searching for formalization query (13.1), in figure 13.5.

$$a - node[tag = "PART"] \quad (13.1)$$

As can be seen, the reading of the results is complicated; they appear stacked and small, and not even the complete sentence can be read, which follows much further on. In addition, although much information saturates the image, not even the morphological information of the Part of Speech appears (beyond the type of Part of Speech itself), which would be helpful. We must also highlight negatively that it does not allow rewriting and manipulation of the trees to highlight the required information in a much more visual and direct way, as seen in other tools later. Finally, it should also be noted that it does not allow automatic downloading of the generated tree conveniently, either as text or as an image.

Another impractical aspect is that it does not allow a mixed search, i.e., qualitative and quantitative, at the same time. In other words, one has to search separately for a phenomenon's total occurrences; on the other hand, one can view such cases. Moreover, it is also important to note that the results are given randomly, without the possibility of obtaining any other type of classification of the examples, based on different criteria that may interest the researcher. The search is also impractical for a Typology/Universals researcher because there is no list of treebanks accessible from the query box, which implies that one has to exit and enter the page every time one wants to change the language or corpus to analyze a fact that also implies a loss of the formula used and, in short, makes the research difficult.

13.4.3 Kontext

Kontext is a tool also hosted by Lindat Clarin (and therefore within the Charles University of Prague environment). It stands out for its simple but functional style. The guide is made in *wiki.corpus.cz*. It could be designed more comprehensively and usefully, but it is visual and useful for the simplest cases.

This tool can filter both by language and by corpus type. In addition to the Universal Dependencies corpora, it has others, such as the European Union. This is because it is not a tool that allows searching for morphosyntactic patterns at all. Dependencies, for example, cannot be analyzed. It is a product more focused on lexical lookup (and lexical context sampling, hence its name). Therefore, its use is highly recommended for research related to the lexicon in Typology or Universals, but it will not be possible to do any other type of research. The tutorial is complete and well-organized.

Positive points

The search is quite simple for a linguist since it allows searching for forms, lemmas, case-sensitive or not, searching by n-grams or groups of words and QCL. This last option is the most useful for linguists more familiar with Computational Linguistics since it allows, for example, a word to be filtered depending on its Part of speech. It is also helpful for diachronic analysis, as it allows adding orthographic exceptions and, for example, not making the distinction between “V” and “U” in Romance languages (in a period when there was hesitation between these). The search result, such as the word “hell”, can be seen in image 13.6.

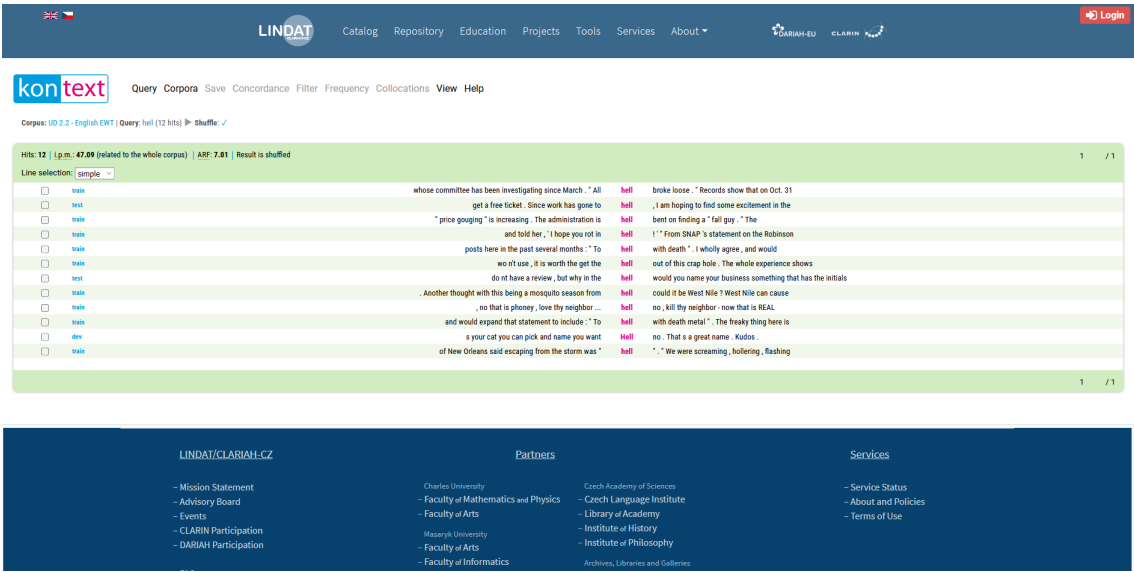


Figure 13.6: General View of “Kontext” Tool

There, you can see in which form the word appears (in the case it has been searched by lemma) and the context in which it appears. You can also see different interesting data, such as how many times it appears (hits), how many times it appears if the corpus had 1 million tokens (i.p.m.), and what the corrected metric concerning the distribution (ARF or Average Reduced Frequency) is. This last metric is a mathematical calculation that allows correcting the distribution of the word in case it appears a lot in a particular place of the corpus and is not homogeneously distributed. Data can be downloaded in different formats (CSV, TXT, XML, XLSX, etc.), which is quite positive and exciting, and even save a URL to the results obtained. Other attractive options are the possibility of filtering results by minimum frequency and creating a ranking with the most frequent forms of lemmas or collocations.

Negative points

The behavior of Kontext is very similar to what is proposed from Sketch Engine (Kilgariff et al., 2014) since both are focused on lexical aspects (no syntax information). Sketch Engine allows you to work with your corpora and not be forced to use already tagged ones, while Kontext does not allow this. Kontext, however, is open access.

In addition to the fact that it only allows lexicon-related searches, Kontext uses each treebank’s own language to make the CQLs (and the corpora are not all labeled homogeneously). This means that, depending on the corpus, the search for a given part of speech can be done with the tag “pos” or “upos”, for example. The whole query syntax is affected depending on the tags of each

annotation system. This makes the search complicated and ineffective. However, as discussed above, it is a good option for lexicon-based research.

13.4.4 Graph Rewriting for NLP - Match

This proposal is hosted at the Inria Center, University of Nancy, as it is a project of Guy Perrier and Bruno Guillaume. It has an online version and a version for use with Python. The Abrorator-Grew-Match tool also aims to facilitate annotating a treebank in UD (or SUD) format in an online, intuitive, and visual way.

Positive points

Grew-Match-match offers an online tutorial that is very complete and useful. It is done in a lesson format, explaining all the possibilities the tool offers, from the simplest to the most complex. In addition, when the user clicks on each feature, a pattern is automatically activated as an example of that search. The last of the steps, related to the clustering of the results, is perhaps one that could be more developed, as it does not express all the possibilities available in the tool (and that one must discover by trying). However, the tutorial teaches effectively and quickly to any linguist, from 0, how to search for word forms, lemmas, Parts of speech, dependency relations, or a combination of these. It also shows how to search for negative values, order or positioning constraints, n-grams, excluding some feature, or a combination of all these things.

Visually, it is a fascinating web where functionality is the priority. The designed patterns allow for highlighting a particular color of what is being searched in the corpus. All the utterances appear to have the context; below, the representation of this in dependencies appears (by arrows). All the information is also displayed, related to the parts of speech and the morphological characteristics of each Part of speech, as can be seen in figure 13.7.

It is also easy to see how the quantitative result can be viewed by simply activating the “count” button. You can customize the amount of information displayed in the image, either by adding or removing “lemma”, “upos”, “xpos”, “features”, “textform”, or “wordform”. Another exciting feature, for example, that was not present in the most similar tool, PML Tree Query, is the ability to sort the results randomly by length or by order of appearance in the corpus. It can also be requested to show more context in case the whole utterance or the metadata is insufficient. It is also possible to download the generated image of the searched information and save the link of the

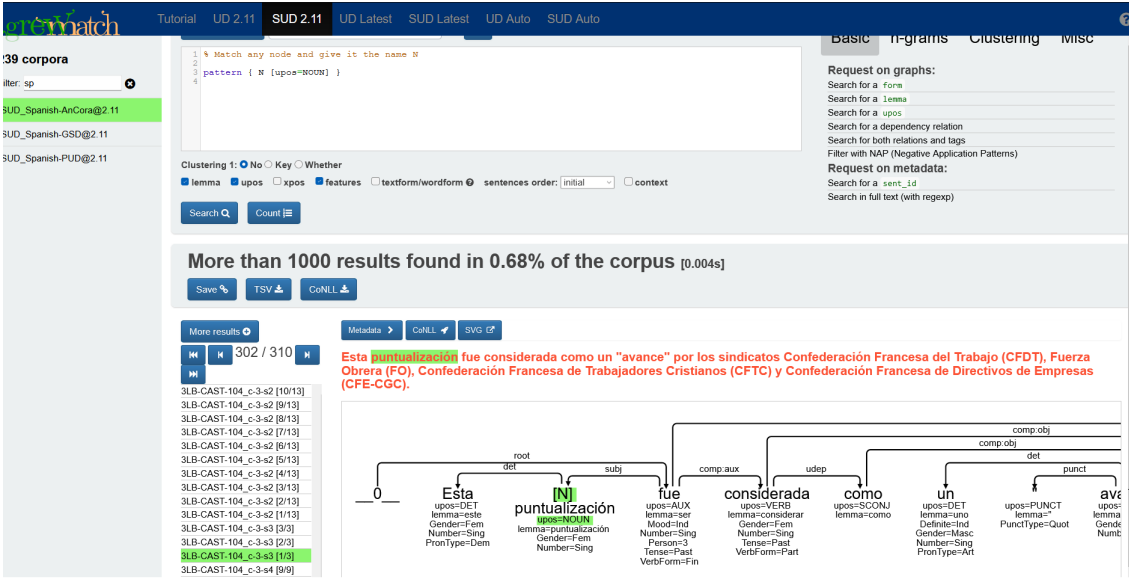


Figure 13.7: General View of “Grew-Match Match” Tool

performed search to activate this search quickly in the future. In this image, the searched items will not only be highlighted in green but also appear with the label the searcher has chosen to give to that particular item. For example, if someone wants nouns to appear with the label NPL instead of NOUN, it is possible to do so because they are looking for plural nouns. Moreover, the tool will automatically change all the labels to whatever the researcher wants.

A search engine is included for the approximately 240 corpora currently in the tool (version 2.11), which allows you to quickly switch between corpora of the same or different languages without losing any information or leaving the page. In addition, it allows switching between the UD annotation system (Nivre et al., 2023) and the SUD annotation system (Gerdes et al., 2022), as each of these may be more beneficial depending on the type of research being carried out.

Hidden options

Other exciting options around the text box are the tables for each corpus, tables in general, clustering by key, and clustering by whether.

The tables of each corpus allow an overview of all results compared and ranked. For example, see figure 13.8. We can see which parts of speech are involved in the direct object dependency relation in this corpus of Spanish. The dependents are placed on the horizontal axis, and the heads are on the vertical axis. This means that, not surprisingly, the most common relationship in Spanish for a direct object is verb as head and noun as dependent. You can see all the cases and the total numbers and click on each case to go to the examples.

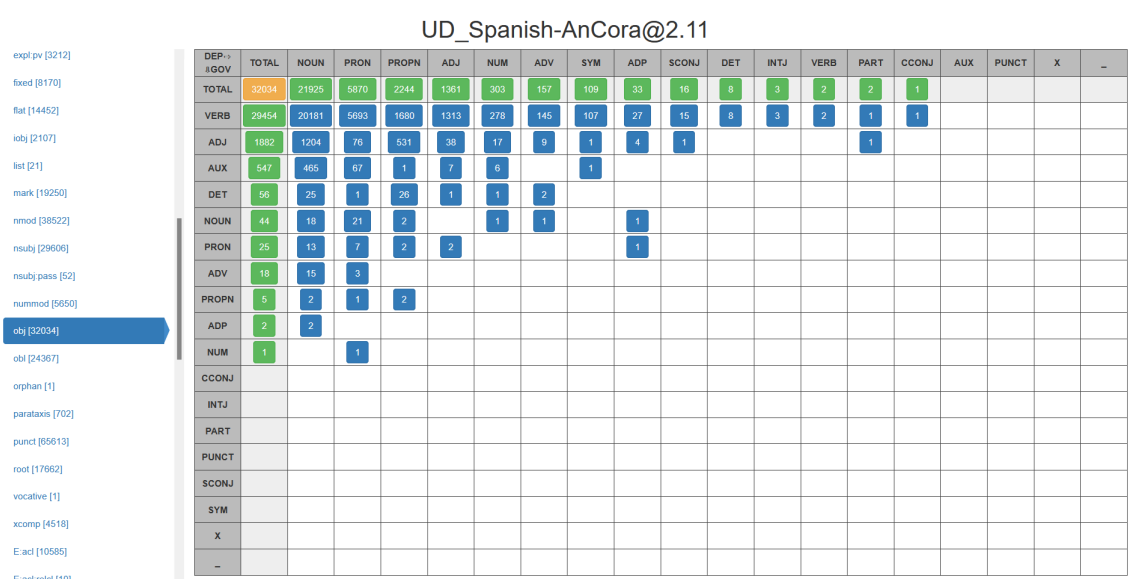


Figure 13.8: Example of the Table on *obj* Dependency’s Part of Speech in Spanish (AnCora)

The tables generally allow you to compare the presence or absence of specific morphological features or dependencies between different languages or corpora (<http://tables.grew.fr/>). For example, see figure 13.9. You can see the presence (and quantity) of the different cases in the different corpora. You can see perfectly which languages have no cases (without any number), which languages have only nominative, accusative, genitive, and locative (Akkadian), or which languages have more (old_east_slavic, for example). You can also click on each number to go to the examples. This can be done with any labels appearing in Universal Dependencies. In addition, you can quickly move the columns to improve the display, add different filters, or prioritize from ascending to descending order of a given column.

The advanced function of creating clusters by key allows us to see the distribution of a particular phenomenon if we classify it by different parameters. These parameters can be any. For example, we can not only ask it to find all the adpositions of a corpus in English and give us the number but also to group them by lemma (i.e., to classify them by which preposition it is), as can be seen in figure 13.10, where the most frequent adposition in English is “of” by far.

You can also do a dichotomous search in which the result is either yes or no. This is useful in cases of word order, for example, where there are only two possibilities (but it applies to any feature, such as whether an adjective is feminine or not). An example of order analysis could be that of the figure 13.11.

In this case, in Catalan, we have asked you to tell us how many cases there are in which the noun precedes the adjective in dependency relations. As we can see, this is the case (yes). However,

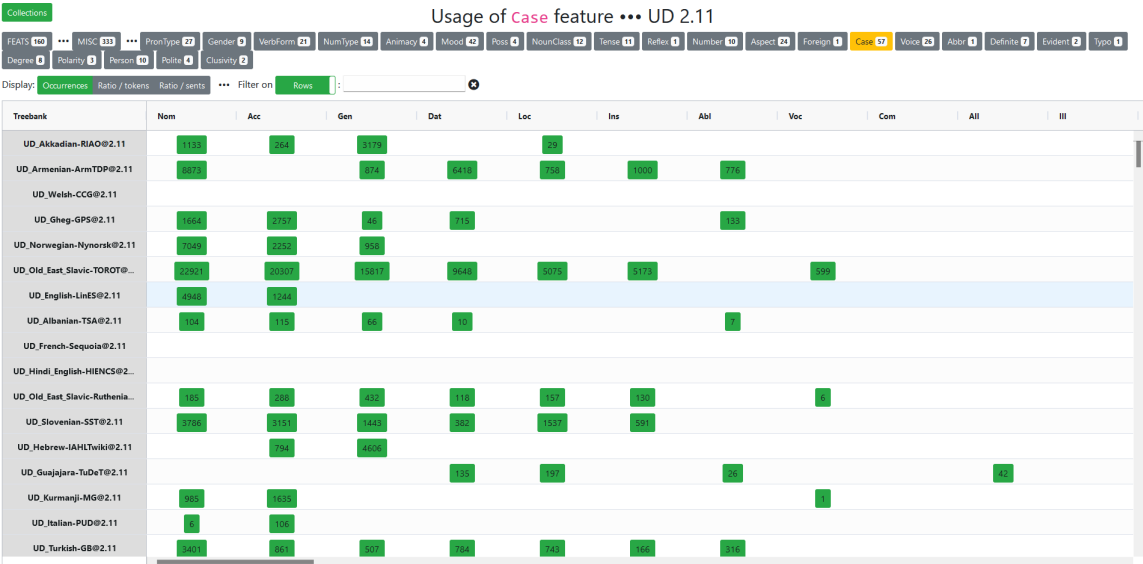


Figure 13.9: Example of Table on Case Types & Frequencies in Languages/Treebanks

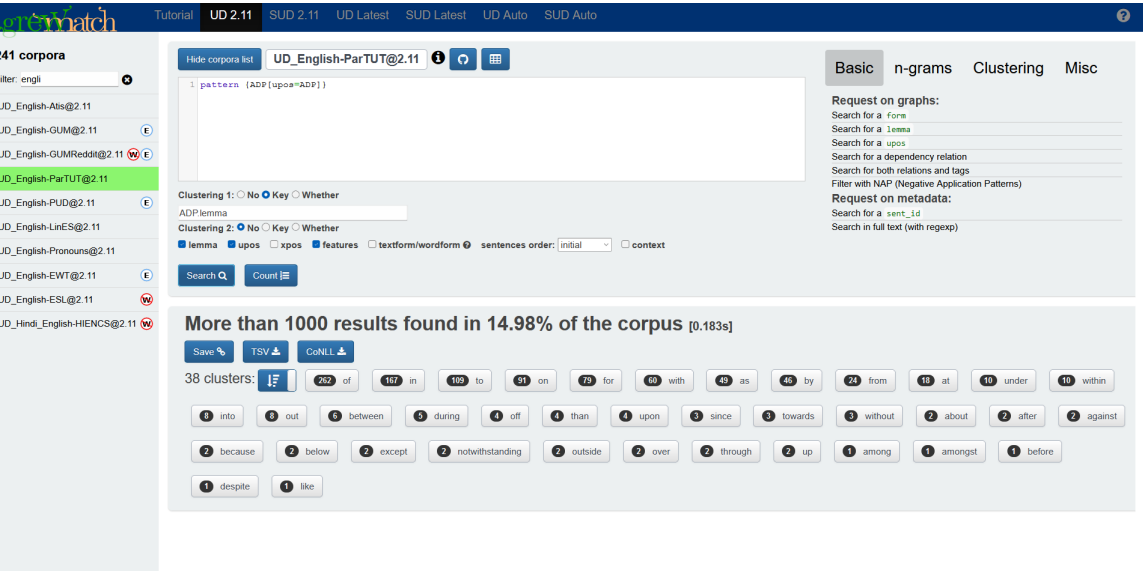


Figure 13.10: Example of Adpositions in an English Treebank, Clustered by Frequency of Lemma

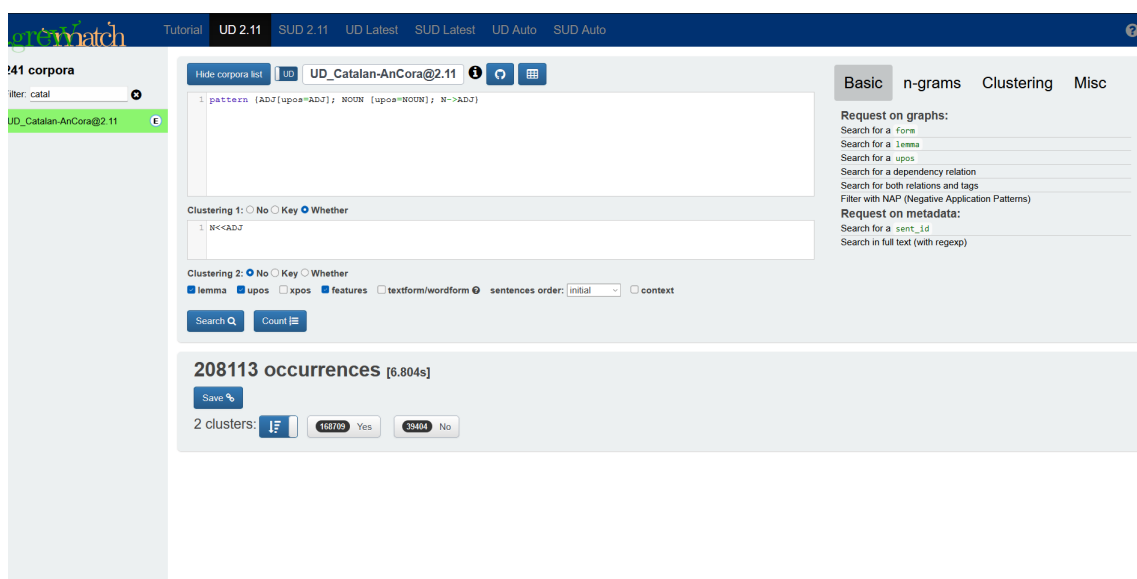


Figure 13.11: Example of Occurrences of Adjectives before and after the Noun Head in Catalan

there are quite a few cases where the answer is no and, therefore, the adjective precedes. If we click “yes” or “no”, we can see the specific cases and examples where this occurs.

Finally, we must highlight the query text box’s great flexibility to customize the pattern as much as we want. You can check adjacencies, specific orders, words with specific characteristics (gender, number, case, polarity, mood, etc.), and several dependency relations at once (for example, look for the pattern of a masculine noun with an adjective as a nominal complement that is the subject of a past tense verb with a pronominal direct complement and a relative clause containing the lemma “yesterday”). This flexibility allows you to be as general or specific as possible in the various analyses.

13.4.5 Infrastructure for the Exploration of Syntax and Semantics

The last resource that allows working with the Universal Dependencies corpora is hosted by the University of Bergen, which we show in figure 13.12.

Positive points

As can be seen, the web page style is relatively sober, focusing primarily on functionality, although it does not make the data easy to read. Many corpora are available, mainly because it is designed to work with LFG (Lexical Functional Grammar) corpora. This means that although the corpora labeled universally, the ones we are interested in, are available, the most updated version

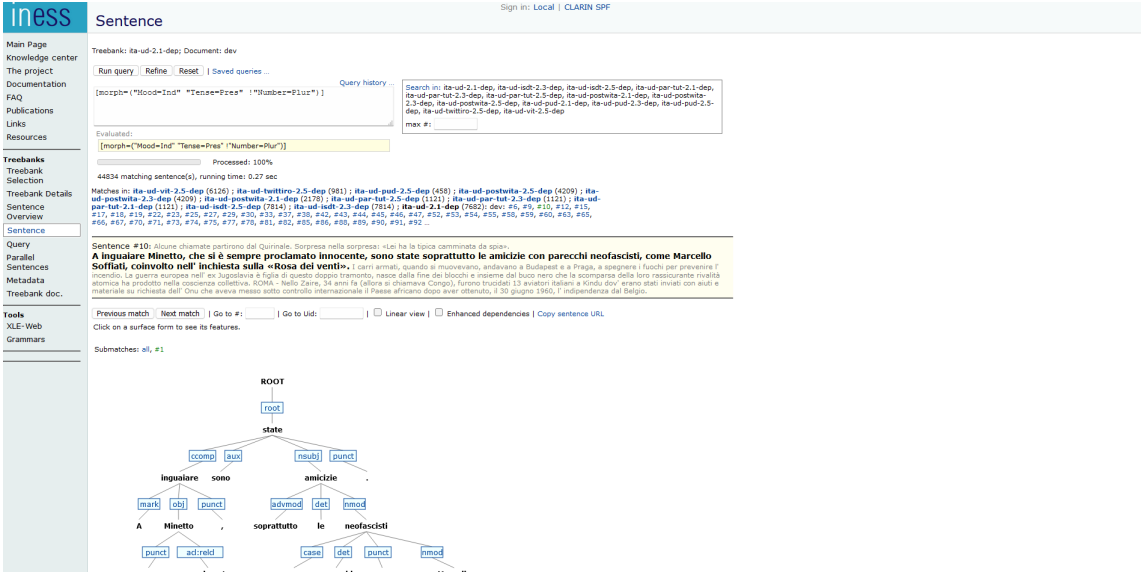


Figure 13.12: General View of “INESS” Tool

of them is not available (nor are all the languages available in UD) since it does not go beyond version 2.8. There is a tutorial available.

These search possibilities are found in Grew-Match, with a broad level of detail in the query. The wide flexibility in its syntax makes possible both very general searches (such as a simple word or lemma) and large combinations of particular dependencies. The nomenclature used, however, is more complex than in Grew-Match. In Grew-Match, the original Universal Dependencies labels are respected, whereas here, everything has its own nomenclature, which must be learned (such as “morph” for “upos”).

One of the positive aspects is that the exact search tells you how many results there are in the rest of the corpus of that language without having to make any additional clicks and in a very visual way. Moreover, you can go to that corpus with a simple click. It is also positive to find the full context of the utterance that has been found and that it stands out from the rest of the text. It is also convenient to see the queries made in the past and be able to rescue and save them.

Negative points

As a negative aspect, we must point out, especially after seeing the functionalities of Grew-Match, that the data reading is not as detailed and complete as it should be. On the one hand, the occurrences are shown in a tree format (but it is not a typical dependency tree, where the original order can be rescued, but a tree corrected to be balanced). However, you can also select the “linear view” option to have arrows that are more in line with the dependency syntax. Therefore, details

about the word order are lost. In addition, the features of each Part of speech are not visible at a glance; they have to be searched for individually.

The guide offered is quite detailed, although not exhaustive. Another problem is that it does not allow us to see and exercise the different patterns that can be made, a didactic process that favors learning. It is not automated; it is simply an old-style guide.

There is also no possibility to highlight anything or change the label name. Concerning the display options, here again, there are no boxes as shown above, nor is there the possibility of making clusters to facilitate the search for a particular element. There is also no possibility to download images or to work with different annotation schemes such as UD and SUD (only UD), which makes it more challenging to represent certain typological studies.

13.4.6 Choice

Once we have reviewed the five tools available for working with Universal Dependencies data, we must choose the most interesting for our research.

The first resource, SETS treebank search, is discarded due to the problems it causes when searching since it is practically out of service.

The third resource, Kontext, is the priority option if we want to carry out lexicon-related searches since it is a tool designed for this purpose (given the convenience of the searches and the different frequency data it offers). This means it will be more advantageous in those situations but does not allow interesting analyses from a morphosyntactic perspective.

We have three options for morphosyntactic analysis: PML Tree Query, Grew-Match-match, and INESS. Having reviewed all three and tested them in some basic searches, we have decided to use Grew-Match-match. This is the most modern and, therefore, incorporates more functional innovations. In addition, it is expected to be more widely monitored and updated. The main reasons for this choice are different functional aspects regarding visualization and general search effectiveness.

From the visualization point of view, we value what we have already shown very positively: the possibility to rewrite the Universal tags as desired (and to highlight them with a different color). The way dependencies are represented is also user-friendly, with an understandable, direct, and complete visualization. For example, it is also important to include (and in a way that does not

load the page with data) all the morphological information of the words, which does not appear the same way in the other resources.

Regarding the effectiveness of general searches, Grew-Match’s tool allows one to prepare very well for the specific searches before facing them. That is, the tables automatically generated by Grew-Match with the characteristics of each treebank compared on a vertical and horizontal axis (with the possibility to click on each case and go to the examples since it automatically creates the query) allow you to see which languages are generally ascribed to the search you want to perform. For example, it is very fast and effective to see which languages have the different categories of “cases” activated with many occurrences. In addition, the columns can be modified, sorted, and dragged, which also facilitates the interpretation of the data. The possibility of making different clusters with the searches performed is also a great visual aid.

Finally, it is also important to mention that this tool is the only one that allows working with the UD annotation scheme and the SUD annotation scheme simultaneously. This is relevant in particular searches, such as the study of adpositions or copulative sentences, where we believe that the SUD proposal (only present in Grew-Match) is much closer to traditional syntax. Also, it is essential, in general, to closer the distance to the syntactic theoretical concepts in Greenberg (which are purely syntactical and not semantic-syntactical, as happens in UD). Another interesting fact is that UD has labels like “xcomp” or “ccomp” that can be applied to direct or indirect objects. So, to be more precise and selective, we find it more interesting to search by SUD labels like “comp:obj”.

13.4.7 Typometrics

Typometrics is a tool created by Kim Gerdes, Sylvain Kahane, and Xinying Chen (Gerdes et al., 2021). It has its website with no institutional or hosting dependencies. It is an exclusive product of its authors.

The main goal of Typometrics is to graphically represent quantitative data extracted from UD corpora of all languages at once. In other words, the aim is to show instantaneously and visually the distribution of the world’s (available) languages about a specific phenomenon (or a combination of two phenomena in a two-dimensional graph).

We have not considered it with the rest of the tools because it does not allow a free search of proposed formalizations. Only concrete and isolated dependencies can be searched, which does not allow the analysis of all Universals.

Metrics

In a few clicks, typological information on a hundred languages can be accessed about their use in the labeled corpora. This information makes it possible to see the behavior of languages in general, but, in addition, by assigning different shapes and colors to typologies or language families, it is possible to see whether this behavior corresponds to a justification of this type. Its easy and intuitive interface allows any linguist without computer knowledge to obtain complete and detailed input with little effort. The data that can be represented are the following:

1. **Menzerath** is studied in the paper of Chen, Gerdes, Kahane, and Courtin (2021). It corresponds to the Menzerath-Altmann law. It is a particular metric based on the calculation of the ratio between the increase of the size of a linguistic construction, resulting in a reduction of the average size of the components that form it. According to the Heavy Constituent Shift, when there is competition in the order of the elements, the shortest in these terms is usually placed closer to its head.
2. **Direction** refers to the positioning of two elements (without indicating their Part of speech) with a dependency. For example, it is the information on whether a language tends to place subjects before or after the verb.
3. **Direction-cfc** refers to the positioning of two elements indicating their Part of speech with a dependency. For example, it is the information on whether a language tends to place pronouns in a dependency relation of the subject with a verb before or after it.
4. **Distance** is the equivalent of “direction”, taking into account the location of the elements in a dependency and the distance between them. The distance increases in relation to the number of words between the two that form a dependency. For example, one could check whether one tends to put more words between the subject and the verb or between the direct object and the verb.
5. **Distance-abs** refers to this same distance relationship without considering the direction in which the dependency goes. In the previous element, placing an element at a great distance before its head will result in a high negative number. In this case, there is no discrimination between positive and negative (or between preceding and postponed elements) since we are only interested in knowing the number of words, regardless of the location.



- ## Example and data reading

At the top left, we have a simple menu with a link to the Typometrics foundational paper (Gerdes et al., 2021) and the tool’s user guide. At the top right, we can switch between UD and SUD annotation schemes. Just below, we find the option to select the creation of a 1D (only one variable) or 2D (mixing two different variables) plot. Just below are the x-axis and y-axis settings. First, select the metric you want to obtain (from the eight explained above) and below the value you want to check. In this case, in the x-Axis, for example, it is NOUN-det-DET, i.e., a name head that receives the dependency of being determined by a determinant. The red dot allows you

to filter n minimum occurrences in case you want to filter to reduce the risk of error. Towards the bottom, we can download the generated graph (yellow cloud) or the generated data (red cloud). In the lower part, there are different configurations to improve the visualization (if we want the names of all the languages to appear, if we want lines crossing the graph to appear or not, etc.).

On the left side, we can see the graph we have generated. In this case, we can see how most languages tend to put the determiner before the noun, and it does not matter if they put the object before the verb or after it; they behave similarly. Only a few languages choose to put it behind without knowing (at least in this graph) what trigger would explain such an exception.

However, this tool also has some limitations or even disadvantages. The main limitation is that only a maximum of 2 linguistic phenomena can be represented simultaneously (in a 2D graph). This means that more complex patterns cannot be queried. In addition, simultaneous occurrences of these two elements cannot be queried either; the search for the two values is independent. Another negative aspect is the fact that there is no possibility to select different parts of speech for the same dependency. On the other hand, it is not possible to work with any morphological information. All this means that the searches that can be made, in the case of Universals, for example, only be the most basic and superficial, without being able to analyze complex constructions.

There are also drawbacks related to the methodology behind the tool. Fundamentally, we can talk about the possible biases in the languages and the mirage of the quantitative approach. Regarding the first problem, we refer to a fact already problematic in Universal Dependencies, which we believe is more accentuated here. This tool provides general data for each language. In the case of multiple corpora, it averages them all. This is very positive. However, many languages only have one corpus, and, moreover, in the case of some specific cases (dead languages, for example), they are usually religious or literary texts (Old Church Slavonic, for example). This means that we are comparing, at the same level, in the same graph, languages with a wide variety of tokens analyzed, with textual variety, etc., with languages with a small number of tokens available, and these are also of a very specific typology that can bias the results. We believe that presenting all the languages in this way, without further information, may cause these biases to be diluted, and we may consider valid and representative results that are mere approximations.

About the mirage of the quantitative approach, we should also mention the philosophy behind this tool, thanks to conversations with the authors (Sylvain Kahane and Kim Gerdes). They do not believe in the classical representation of Typology by labels. They believe this classical conception limits the study and produces many errors of perception (because labels are simplifications

of reality, as we all can agree). Therefore, in this proposal, everything revolves around purely quantitative data. The authors' interest lies in seeing what graphical or mathematical distribution can be produced once the points relating to each language are placed.

The main functions or distributions expected to be found are exemplified in the paper on creating Typometrics (Gerdes et al., 2021). Some can be the Z-distribution or the almond distribution (for combinations of two typological features in a 2D graph). In a Z-distribution pattern, languages of a given type and another would be shown at the two extremes of a graph about another linguistic feature. On the other hand, languages that would fit more into the “non-dominant order” type (even though they do not use categories) may also present central values in the other linguistic features analyzed, creating a Z-like figure. An example of this distribution in our study could be the one shown in figure 13.14 regarding Universal 4, in the “direction-cfc VERB-comp:obj-NOUN ADP-comp:obj-PRON” part.

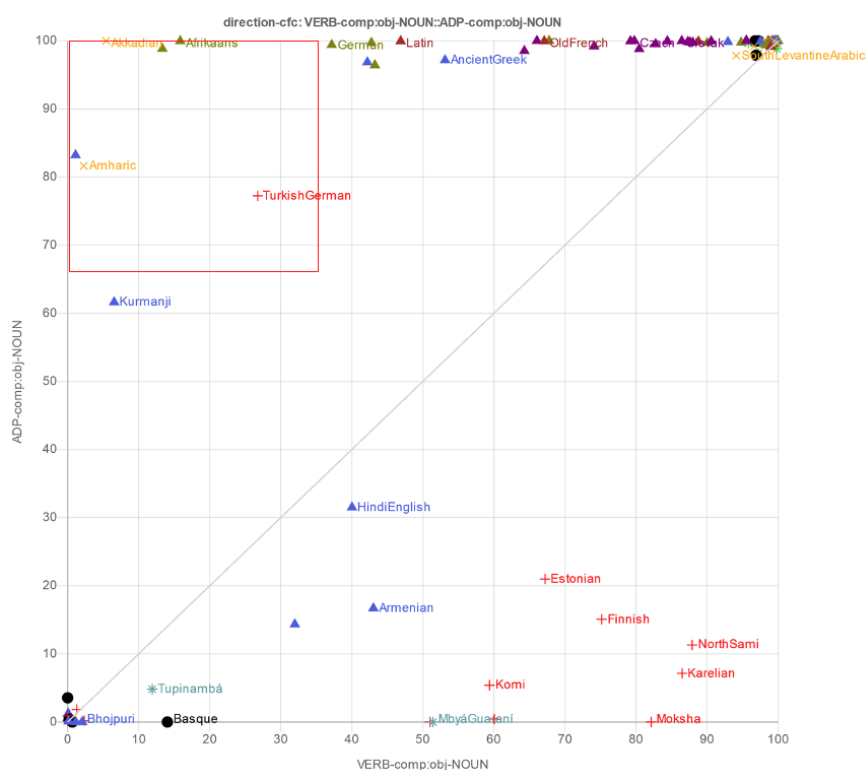


Figure 13.14: Example of Results in Typometrics Capturing the Relation between Nominal Objects and Adpositions

Another key element for the authors has to do with the triangle drawn, thanks to the diagonal in the graph. This marks the equality in the quantity of element X on the graph and element Y. For example, if a language is right on that line, it will present the same direct complements as subjects

after the verb (if that is what was being investigated). Returning to what we have labeled as the mirage of the quantitative approach, we believe that we risk giving too much importance to data that are no more than a reflection of a given corpus, in a given place, for a given language. In other words, it may be misleading to state that Catalan is a language that places more adjectives before the noun than Spanish because we have a result of 45% in one and 41% in the other. Perhaps, with other corpora, the data would be slightly different. Therefore, we believe that these data are relevant to indicate trends or to be able to place a language in a specific group but not to speak in exact terms that do not reflect more than a specific example (hence the interest in our fuzzy modeling proposal). Obviously, the other problems that exist in UD (such as the lack of typological variety in the sampling, for example) are also present in Typometrics. However, these are not highlighted again because they have little room for improvement in Typometrics.

On balance, we believe that the benefits offered by Typometrics outweigh the disadvantages. We also believe presenting results as visually as possible is very interesting. Furthermore, it may also be essential to present studies based on this novel tool so that it can become known in the language Typology community, demonstrating how it works for linguists who were probably unaware of it. Therefore, whenever possible, we will present data analyzed using Typometrics (without this implying that they are no longer formalized and searched in Grew-Match). We believe this extra output can be very interesting. However, we can only do it in very few cases (the Universal requires a very shallow analysis and is very general).

In order to compare the possibilities offered by Typometrics with the rest of the tools, we present, as a summary, table 13.1.

Feature	SETS	PML-TQ	KonText	GREW	INESS	Typo-metrics
Useful for syntax	-	+	-	+	+	+
Allows complex queries	-	+	-	+	+	-
Quantitative visual output	-	-	-	-	-	+
Has a full tutorial	-	+	+	+	+	+
Morphological information displayed	-	-	-	+	-	-
Visual and organized results	-	-	+	+	-	+
Simultaneous quantitative and qualitative results	-	-	+	+	-	-

Table 13.1: Summary of the Characteristics of the Analyzed Tools

UNIVERSITAT ROVIRA I VIRGILI

GLU - A GRADIENT MODEL FOR LANGUAGE UNIVERSALS. A COMPUTATIONAL REVISITATION OF GREENBERG UNIVERSALS

Antoni Brosa Rodríguez

14

Fuzzy Logic

Contents

14.1	Gradience and Typology	196
14.2	Gradience and Universals	197
14.3	The Sorites Paradox	198
14.4	Fuzzy Natural Logic Concepts	199

In linguistics, as in many other disciplines, there are essentially two views of the nature of phenomena. The perspective behind discrete logic understands and models the object of study as a Boolean manifestation, i.e., with two possible values. The perspective behind Fuzzy Logic understands and models the object of study as a continuum.

We understand that the study of Universals is no stranger to this dichotomy. In the case of deep Universals, those proposed within the logic of generative grammar, there is a clear commitment to the Boolean functioning of language. This is also affirmed by Torrens-Urrutia (2019:49), for example:

This attitude towards gradience comes from the theoretical distinction of competence and performance, which rules generative frameworks. Because of this distinction, they used to avoid empirical data since they conceived grammar as a perfect mathematical framework in which everything is either-or-that in the deep structures, placing the vagueness, randomness, and empirical data as things that might occur in the performance.

It is coherent to renounce the empirical data of Chomskyan Universals insofar as that would be to resort to performance. Consequently, at the level of competence, we only find the possibility of having Universals or not having them.

14.1 Gradience and Typology

In the case of Greenbergian Universals, we are analyzing empirical data, leading us to the formulated theory. This, in theory, should place us in an environment that takes vagueness into account and avoids a classification, as Torrens-Urrutia (2019) comments, of “either-or-that”. However, we consider that this has not been the case.

While it is true that there is a consensus in Linguistic Typology about the acceptance of exceptions, we consider that their fit has not been adequately resolved. We believe the current classification of typological Universals is somewhere between a discrete and a diffuse understanding of language. It is this fact that might cause difficulty fitting the current model.

There is a shift from a duality between Universals and non-Universals to three elements: absolute, statistical, and non-Universals. However, no model exists to capture the existing gradation in the universality of rules.

The classical classificatory approach distinguishes between absolute and statistical Universals. This separation is sharp. If a rule is present in all languages, it is an absolute Universal. If it has one or more exceptions, it is a statistical Universal, without entering into any kind of gradation. The same consideration applies to the distinction between statistical Universals and non-Universals.

This means that if we adopt a discrete model, we would see graphically how the crisp set would have a shape like a square. There is no gradation (increasing or decreasing) in the change with the next set. In a practical example, if the set of young has its end before 30, it means that any element with the value 29 would be considered young for all purposes and without nuance. In contrast, when it becomes a value of 30, it ceases to be young altogether.

A clear example of this functioning in another aspect of Typology is the use of labels to locate languages. In these, there is an inherent simplification of linguistic types. This causes everything to be reduced to discrete categories. Languages are classified by exhibiting one characteristic or its opposite without reflecting further details. A discrete view assigns an order of adposition to languages: prepositional or postpositional languages. From the Fuzzy Logic, the concepts of preposition and postposition would be maintained, but these would be part of the two extremes of a scale. Being able to reflect the location of a language at a non-extreme place on the scale (e.g., a prepositional language with some postpositions) may allow a better understanding of its behavior in other linguistic aspects.

The discrete approach, helpful in favoring comparisons, can complicate comparisons in other respects. A self-explanatory example of this is the inclusion within the SVO type of languages with remarkable flexibility, such as Russian, and, at the same time, rigid languages, such as English. In other parameters, their behavior will not be exactly the same since, although both have this order as a basis, it is not equally frequent.

This type of criticism is beginning to be made in recent years, although one can already guess this problem in previous attempts (Greenberg, 1978b; Krámský, 1959). Levshina et al. (2021) clearly shows the advantages of a paradigm shift by adopting a gradient approach to the typological features of languages. She points out that it was impossible to carry out such studies so far due to the lack of technical possibilities (digitized corpora). It seems much more precise and exciting to understand the distinction between OV and VO as a continuum and not as a question of 0 and 1, for example.

14.2 Gradience and Universals

Ponti et al. (2019) also point out that many of the categories in natural language are gradient, even though we conveniently treat them as having a discrete character. Therefore, we dare to propose that the concept of Universality is also gradient. We believe that the conception of Universals as gradient was already, in a way, indirectly assumed by researchers.

The creation of statistical Universals can be understood as something that aims to provide gradation to differentiate between two possibilities: non-Universal and Universal. This use of statistical, in a translation of the non-Universal as a value 0 and the Universal as a value 1, would place the rules analyzed at a point close to 1, without being 1. In a certain way, we understand that de facto, when scholars on Universals equate the value of a statistical Universal and an absolute Universal, they are understanding Universality in a matter of degrees and assume that the difference between absolute Universals (1 or maximum value of the scale) and statistical Universals (almost 1 or close to the maximum value of the scale) is not significant. However, we have not found (up to our knowledge) any theoretical reflection about this fact clearly and explicitly in the literature, as proposed in Torrens-Urrutia et al. (2021). In other words, Universality can be understood as a continuum, in which the investigated element can be located closer to 1 the more languages of the set share this element.

We do not consider it relevant to establish a precise cut-off point between absolute and statistical Universals and neither between statistical Universals nor non-Universals. It must be remembered that the formulations of Universals are highly dependent on the sampling. If this precise cut-off point were maintained, adding or deleting a language could completely change the cataloging of the Universal. However, as typologists point out, the important thing is to detect the general behavior of the languages of a linguistic type without assessing whether this characteristic presents any counterexample.

14.3 The Sorites Paradox

Therefore, we have considered it relevant to investigate what other proposals have been made to solve similar problems. Égré and Klinedinst (2011:4) show which is the best way to deal with this kind of problem:

From a semantic point of view, vagueness raises the question of which framework might be the most suitable to represent the absence of sharp boundaries for most of our predicates. Classical logic is well suited to deal with sharp predicates. However, it offers no direct handle on vagueness: it was developed precisely to deal with the language of exact sciences and determinate concepts.

Since ancient Greece, problems have been presented that reflect the limitations of the Boolean model in understanding some aspects of the world. A representative example of this is known as the Sorites Paradox. This deals with the inconsistency behind some concepts, such as the heap of sand. We can agree that two or three grains of sand do not form a heap of sand (Égré & Klinedinst, 2011). Likewise, a million grains of sand do form a heap. If one grain is added to the two or three grains, it will still not be a heap. It will still be a heap if you subtract one grain from the million grains. If you keep repeating this action, there will come a point where there will be an overlap between the two poles. However, we will be unable to determine how many grains of sand are necessary to speak of a heap.

In other situations, we can find this same problem in many adjectives such as, for example, bald (Cook, 2002). In some cases, it is difficult to determine when a person is considered bald or not. This, from the logical point of view, we can reflect it as does Égré and Klinedinst (2011:2): “The abstract formulation of a sorites paradox is given by the mathematical principle of induction, which states that if a predicate P holds for 0, and if it holds for $n+1$ whenever it holds for n , then the predicate holds for every n ”.

Outside linguistics, this idea is at the basis of some industrial innovations such as autonomous cars or air conditioners.

Fuzzy logic is the discipline that deals with these issues and is opposed to the boolean logic we have seen above. Novák, Perfilieva, and Mockor (1999) define it as a type of many-valued logic. That is, it is understood that the truth value of any element, instead of being 0 or 1, is between 0 and 1 (Smith, 2008). Each element will have a degree of belonging to this group (Zadeh, 1965:338).

In short, the perception we have presented applied to situations such as the heap of sand or the adjective bald can be applied to the concept of universality. We start from this conception based on Fuzzy Logic and take one of its models to better approach the phenomenon of Universals.

14.4 Fuzzy Natural Logic Concepts

The concept we mentioned in Levshina et al. (2021), gradience, is a common label in linguistics that can encompass things that, in Fuzzy Logic, are known as vagueness and uncertainty. Vagueness is encountered when it is necessary to group into X some objects $(x, y, \dots)$ that exhibit the property φ (Torrens-Urrutia, 2019):

$$X = \{x \mid x \text{ has the property } \varphi\} \quad (14.1)$$

Instead, uncertainty has to do with the lack of knowledge about the occurrence of a given object (Torrens-Urrutia, 2019) in a context. This issue related to probability is not the one we are working with here in a direct way, and, therefore, we try to deal with the vagueness of the concept of Universal.

Most of the works that relate linguistics with Fuzzy Logic are in semantics (Lakoff, 1973; Novak, 2021; Novák & Murinová, 2019; Novák, Murinová, & Boffa, 2020). However, in recent years, we also find some proposals to assess the grammaticality of structures (Torrens-Urrutia, 2019). We believe this approach can be applied to many other unresolved linguistic issues, as could be the mutual intelligibility of varieties (Hammarström, 2008). In our particular case, we believe that we can take advantage of the models proposed within what is known as Fuzzy Natural Logic (FNL) for these different issues and adapt them to be able to model our object of study.

What may be interesting for the reader is to know the concepts behind FNL, which are the ones that will appear in the model used. Torrens-Urrutia, Jiménez-López, and Novák (2022) summarize them perfectly in 6 fundamental notions we will discuss:

1. It is based on the notion of a Fuzzy set. “Given an arbitrary set V (a universe), by a fuzzy set, we understand a function $A : V \Rightarrow [0, 1]$ where the elements $a \in [0, 1]$ are interpreted as truth values” (Torrens-Urrutia et al., 2022:66).

That is, all the elements that we classify belong to the same universe. Within this universe, each element has a value between 0 and 1, which is its degree of belonging to this fuzzy set.

2. We must refer to Lakoff’s hypothesis on Universal meaning. This principle lies in the understanding, from a general point of view, among any human being, no matter what language he speaks. The values of a concept are present anywhere in the world. Torrens-Urrutia et al. (2022) offer the example that the meaning of *tall* is Universal.

The Universality of the meaning of *tall* is present in any culture or language. Although the label used in each language is different and the value assigned to *tall* may vary between the societies of China and the Netherlands when applied to a man, for example, the essence of *tall* is the same. That is, *tall* means a high value in height. This shared common framework makes the meaning Universal and can be applied without problem.

3. This application is made to evaluative linguistic expressions (Novák, 2008a; Novák et al., 2016; Torrens-Urrutia et al., 2022). These are natural language expressions used to evaluate any object or entity.

These must contain two independent words that act as ends of the scale. Within that scale, gradation is allowed. For example, we could have the word *high* at one end as the equivalent of 1 and the word *low* as the equivalent of 0.

4. The possible world refers to “a particular context in which a linguistic expression is used” (Torrens-Urrutia et al., 2022:66).

These possible worlds allow us to represent the variability in interpreting the fuzzy values of the variables in a specific reality or context.

5. We understand by intension “a property denoted by a specific world or expression. It is independent of a specific possible world (context) and does not change when the latter is changed” (Torrens-Urrutia et al., 2022:66).

In a fuzzy set, the intension of an element is expressed by a fuzzy value representing the degree of belonging of the element to the set. This value can vary from 0 to 1, where 0 indicates that the element does not belong to the set, and 1 indicates that the element belongs completely to the set. Intermediate values express partial degrees of membership. Intension allows quantifying and working with degrees of certainty or membership instead of the classical dichotomy of membership or non-membership.

6. We understand extension as the referent: “it depends on the particular context of use and changes when the possible world has changed” (Torrens-Urrutia et al., 2022:66).

That is, the context in which the analysis is being applied determines the extent of the analysis.

To see the interrelation between these last three elements more clearly, Torrens-Urrutia et al. (2022) offer a very illustrative example. If the context is “driving a car”, the intension “very long distance in Europe” can lead to an extension of 1000km. If the context is “walking on foot”, the intension “very long distance in Europe” can lead to an extension of 5km.

That is, it can be seen, as we have commented above, that the intension (very long distance in Europe) does not vary in any case, it can be considered a set equivalent to “very short distance in Europe” and each of the elements analyzed has a degree of belonging to this set. In the materialization of the trichotomous expression we use, its design is defined by expert judgment. However, this coincides with a representation of fully balanced sets.

The possible world is equivalent to the context selected for modeling and can correspond to any situation. In the example, two possible situations are shown: “driving a car” and “walking on foot.”

The two situations described condition the result of the extension in a different way. This determined extension can be 1000km in the first case and 5km in the second case.

The specific Fuzzy Logic model we have selected to model is based on a Trichotomous Expression. Its suitability comes from its successful application in different linguistic aspects in Novák (2008a); Novák et al. (2016) and Torrens-Urrutia, Novák, and Jiménez-López (2022). Parallelisms can be observed between the modeling of syntactic or semantic aspects that can apply to the study of Universals. It is explained and developed in the following section.

We must remember that this proposed interdisciplinary study is not within the scope of Fuzzy Logic but is simply modeled with a Fuzzy Logic-based model. As Cook (2002:226) points out, the

aim is not to provide descriptions of linguistic phenomena but to model the vagueness of these elements. Therefore, we do not consider necessary a detailed mathematical analysis of the model used or an in-depth commentary on the mathematical terminology, since this is not the purpose of the thesis. If the reader wants to expand on these aspects, we recommend reading Novák (2008a, 2008b); Novak (2021); Novák et al. (2020, 2016).

Part IV

A Gradient Language Universals

Model - GLU

UNIVERSITAT ROVIRA I VIRGILI

GLU - A GRADIENT MODEL FOR LANGUAGE UNIVERSALS. A COMPUTATIONAL REVISITATION OF GREENBERG UNIVERSALS

Antoni Brosa Rodríguez

15

Fuzzy Logic Models

Contents

15.1	Five Triangular Fuzzy Sets	205
15.2	Five Trapezoidal Fuzzy Sets	207
15.3	From Five to Three Fuzzy Sets	208

We have opposed the models that work with crisp and fuzzy sets in the theoretical framework and shown the main features. We have argued our opposition to the prevailing model in the conception of Universals (in crisp sets) and have proposed a fuzzy set model for Universality in languages. However, there is not only one model in Fuzzy Logic. In this chapter, we show the discipline's most widely used and well-established options and evaluate them to choose the most helpful representation for the case of Universals.

15.1 Five Triangular Fuzzy Sets

A frequent and common model in Fuzzy Logic is the one representing five triangular sets. It is easy to find this representation in various industrial applications where the concept of maximum membership degree to the fuzzy set is linked to a single central point. For example, we consider this type of membership function to be very useful in using air conditioners. Working with the temperature and operation of such devices can be perfectly represented as described. This model is used by Gajovic et al. (2018) in the figure 15.1.

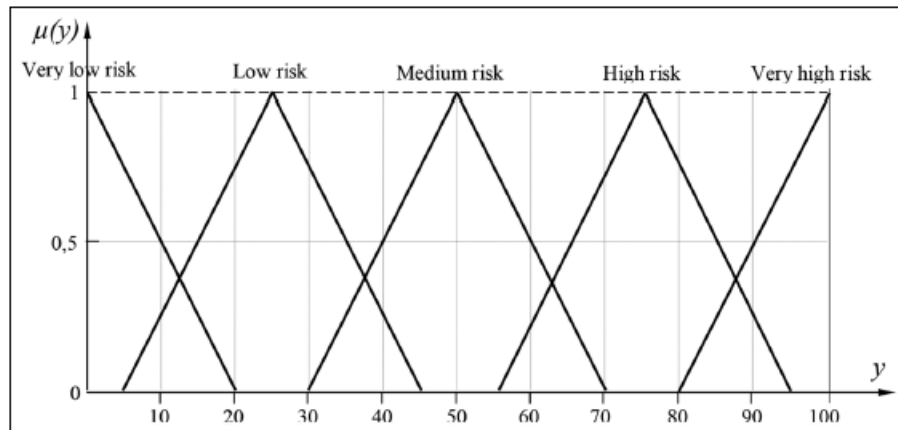


Figure 15.1: Model with 5 Triangular Fuzzy Sets in Gajovic et al. (2018)

This is characterized by:

- **Having fuzzy sets.** The sets present are fuzzy. This feature must be present in any Fuzzy Logic proposal. This means that the change or transition between one set and the next is not crisp or sharp but that there is a gradation between them.

In this case, the sets are ascribed in a triangular figure. This allows the opposite of a square (which would be crisp): the degree of set membership is gradually modified as one moves up or down the scale. Changing the position of an element on the x-axis has implications for the value assigned on the y-axis, which is linked to the set membership value.

- **There is no overlap between non-contiguous fuzzy sets.** The gradation of the set change mentioned above is not arbitrary. There is never any contact or overlap, not even minimal, between two non-contiguous sets. The model allows us to speak of a possible duality of an element if it has a degree X of low and medium, for example. It is not logically adequate to speak of an element being able to have a degree X of low and high membership simultaneously.
- **The fuzzy sets are 5.** The fuzzy sets, in this case, are very low, low, medium, high, and very high.
- **The fuzzy sets are triangular.** This causes the core, that is, the part of the set where we have a set membership value of 1, to be very small (one specific point). The direct consequence is that few elements can be placed in this area of maximum prototypical value or 1 as a membership degree to the fuzzy set.

- **There is no overlapping between the end of the support¹ of a set and the core of another set.** The gradation also follows a specific logic in combining the vertical and horizontal axes. It is not logically adequate that an element Y can have a degree X of low and be an element prototypically with a maximum value (1) of medium. There is no overlap between the end of support of a set and the maximum value (1) of belonging to the next set.

15.2 Five Trapezoidal Fuzzy Sets

The model shown in the previous section has not convinced some researchers because of the use of triangular structures. In some disciplines, such as linguistics, having a straightforward way of marking prototypicality at its highest level is interesting. This has led to the search for other forms, mainly trapezoidal and Gaussian. This change can be seen in the model used in Moreno Cárdenas et al. (2020) in the figure 15.2. The difference between a triangular and a trapezoidal fuzzy set lies in its core configuration. In the case of a triangular set, the set design and its membership degree 1 will pivot around a particular number as the core. In the case of a trapezoidal set, its core is a range. In a Gaussian set, the transition between sets is even more gradual.

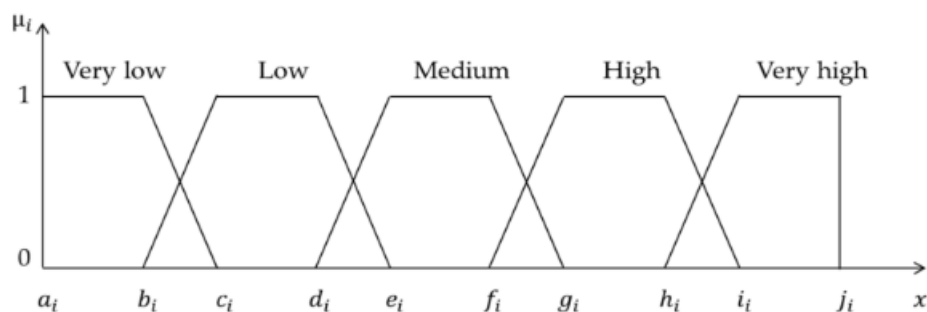


Figure 15.2: Model with 5 Trapezoidal Fuzzy Sets in Moreno Cárdenas et al. (2020)

In this case, we can see how the characteristics shown in the original model are maintained, except for the aforementioned change from the triangular membership function to the trapezoidal membership function. These fuzzy sets are five again: very low, low, medium, high, and very high. In the design, there is no overlap between sets 1 and 3, 2 and 4, or 3 and 5. There is also no overlap between the end of a set and the beginning of the maximum degree of the membership degree of the contiguous set.

¹We understand by *support* all the included zone with a value higher than 0 within a set. When we refer to the end of the support, we mean the place where the set has a value of 0 in the vertical axis or its membership degree.

In the case of the cited article of Moreno Cárdenas et al. (2020), the model is applied to a particular case: the study of coffee mucilage wastes as a possible renewable energy source. The article demonstrates the suitability of the fuzzy-logic-based model compared to other models that offer lower accuracy results.

15.3 From Five to Three Fuzzy Sets

The models shown above are efficient and suitable for concrete and complex situations. However, Novák (2008a) and Novák et al. (2016) consider the division into the five sets problematic for the work done.

We recall that the two examples above use the sets very low, low, medium, high, and very high. Novák (2008a) and Novák et al. (2016) point to a problem in the distribution of the sets from a logical point of view. The problem lies at the extremes of the scale. We are looking at the existence of two separate sets for high and very high, for example. This distinction has implications that can be problematic:

- Although there is a transition zone, as in any set, there are points of non-coincidence. This assumes that we are considering that there are elements that have a degree of belonging to the very high set and that do not belong to the high set. That is, we can infer that something very high is not high.
- The maximum membership degree of the very high set coincides with the 0 membership degree of the high set. This implies that we are indicating that when a given element is a representative of something very high at its maximum level, it is not high at all.

We also consider that if an element Y belongs to the very high set, it should also be understood as belonging to the high group at its highest and prototypical values. We do not understand these two sets as mutually exclusive. This exclusion can be seen among other sets. For example, if an element is prototypically high, we understand it to have no degree of belonging to the medium or low set. Therefore, we believe that not all sets exhibit the same characteristics from a logical point of view among them. In other words, if an element has a membership degree of 1 in the very high set, it should also have a membership degree of 1 in terms of high. If an element is very high, we understand it is also clearly high. On the other hand, an element with a membership degree of 1 in the high set should not have a membership degree of 1 in terms of medium.

Novák (2008a) and Novák et al. (2016) do not only focus on pointing out this situation but also propose a solution. The change is based on merging the low and very low sets and high and very high. In this way, we put an end to this apparent contradiction.

Such a merge may lead to a loss of a meaningful distinction between the sets low and very low, for example. However, it should be noted that the division into three sets is not incompatible with the detail of such aspects (such as very low or very high) if the researcher considers such sets necessary or exciting for the study. In such a case, what should be done is to add the distinction between low and very low within the low set and not as a separate set, as it appears in the figure 15.3 of Torrens-Urrutia, Novák, and Jiménez-López (2022).

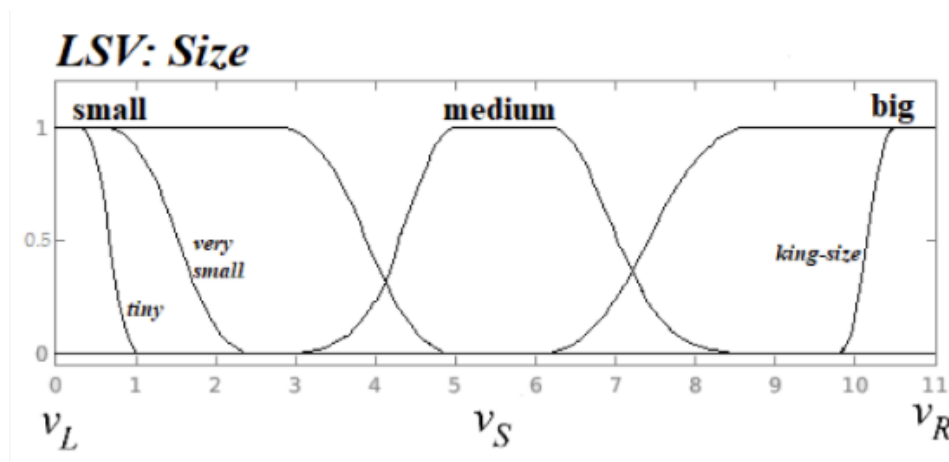


Figure 15.3: Trichotomous Expression Proposed by Torrens-Urrutia, Novák, and Jiménez-López (2022)

In this example, the consistency and flexibility of the proposal can be observed. Consistency can be seen in using three sets, all of which are now equally exclusive. Flexibility is observed in the possibility that the researcher can reflect smaller and more precise distinctions (such as tiny or very small) if necessary, but always within the largest set it encompasses. In the case of very small, for example, there is a correspondence between the distinction made in this proposal of 3 sets and the previous division into 5 sets in terms of their proportional location. However, in this case, the membership degree of degree 1 of small does not vary when very small ends or begins. That is, any element within very small is within the set of small.

We find applications of this proposal, especially in the field of linguistics (Novák, 2008a; Novák et al., 2016; Torrens-Urrutia et al., 2022; Torrens-Urrutia, Novák, & Jiménez-López, 2022):

- Studies on quantifiers. In particular, work is done on intermediate quantifiers, such as most, a lot of, many, a few, a great deal of, a large part of, a small part of, etc. The small-medium-big triplet is used for this study.
- Studies on evaluative expressions. We find proposals to study expressions such as “The tower of Burj Khalifa in Dubai is extremely high”. The triplet small-medium-high is used. This case is an excellent example of how these further precisions can be added within the sets, something that happens with extremely high. The authors represent such a distinction, but always within the high set. In such a study, one can see the implications of changing the context (from Dubai to Europe, for example). Another interesting aspect of this research is the correlation between positive and negative degrees from 0 to 1.

16

Fuzzy Logic Model for Language Universals

Contents

16.1	Description of the Model	212
16.2	Example of Application to the Universal 13	214
16.3	Comparison of Traditional Labels and the GLU Model	217
16.4	Benefits of the GLU Model	219
16.5	On Possible Objections to the GLU Model	222

It is this last model that we have presented (Novák, 2008a; Novák et al., 2016) that we take as a reference for our proposal on Language Universals. On the one hand, we consider that the characteristics we discussed best fit with the proposed study. It is the model that allows a more natural transition between sets. It allows a clear opposition between elements with a maximum degree of belonging to the set and those without an understandable range. The existence of three fundamental sets makes it possible not to fall into the contradiction of opposing two sets that have a clear link semantically and logically. On the other hand, it has already been implemented in other linguistic works. So we expect it to be easier to apply.

16.1 Description of the Model

The model we present is called the Gradient Language Universals (GLU) Model. The name refers directly to the change in the conception of Universal we propose. We follow the proposal known as trichotomous expression, in which we have three sets corresponding to a triplet of any evaluative expression. We have already seen several examples, such as small-medium-big. We consider that we can work with universality perfectly well with the low-medium-high evaluative expression. This has been used in examples we have shown before, and we consider that it combines very well with our proposal. Thus, we understand that a rule (or a Universal) can have different degrees of universality and, therefore, be a Low-Universal, a Mid-Universal, or a High-Universal. This trichotomous division is coherent, not only with the set reduction proposal of Novák et al. (2016), but it also coincides with the division made by Smith (2008) into the positive set, negative set, and middle set.

The scale shows three fuzzy sets. Our main objective is to identify the rules with a high degree of universality and, therefore, are shared by most of the world's languages. We do not consider making further distinctions within the High-Universal group relevant, mainly because one of our purposes is to reduce the existing terminology and simplify the classification as much as possible. However, if any researcher considers it relevant, modifying the configuration to include such distinctions is possible. It should be remembered that we are not talking about creating more sets in the model used but rather about establishing distinctions within the same set.

The model we use can be seen in figure 16.1 and is based on what we have proposed in Torrens-Urrutia et al. (2021).

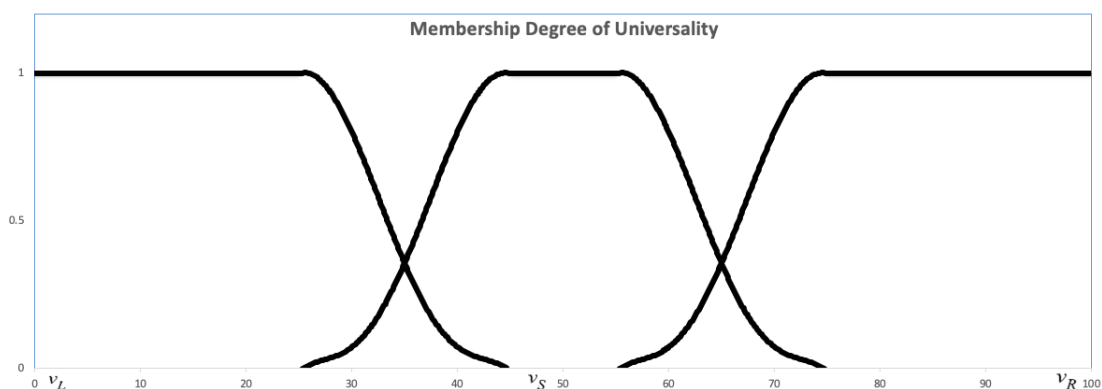


Figure 16.1: Gradient Language Universals Model

We recall that we have three sets to place the trichotomous expression low-medium-high. By convention and coherence, we respect this order and, therefore, we assign the left set to low (V_L), the central set (V_S) to medium, and the right set (V_R) to high.

The sets maintain the proportions of all the sets seen above. Although these may be based on expert judgments or complex mathematical formulae, the result is a proportional and equal scale division. In other words, the guidelines of the Novák et al. (2016) model are followed. The guidelines of the models presented above are also followed if the proportions of low and very low and high and very high have been merged.

There is a good fit between the model we present and the claims made in quantitative Typology. Levshina et al. (2021) and Gries (2013) provide a model for dealing with dominance. This concept is central to Greenberg and attempts to show when a linguistic feature is dominant in a language. They present a criterion that, although it may be discrete, coincides with the model's cut-off points. They propose that a linguistic type is dominant in a language if it appears twice as often or more than the second most attested option. In the case of a dichotomy between two basic linguistic types, such as object-verb languages and verb-object languages, we can perfectly well use the model we propose. The left set could be for OV languages, the middle set for the non-dominant order, and the right set for VO languages. Therefore, we believe that the model we propose for classifying Universals in language can be useful in Typology for more things, such as the representation in terms of Fuzzy Logic of linguistic types.

This ratio also takes into account other factors which are essential in the models already analyzed. The three sets are linked by superimposed Gaussian curves that favor the natural transition between fuzzy sets. The core of the sets corresponds to each set's usual range, allowing the opposition between prototypical and non-prototypical structures to be established. There is no overlap in the support of the low and high sets. There is no overlap between a set's end and the next set's maximum membership degree.

On the y-axis, we find the information related to the degree of belonging of an element to the sets, which ranges from 0 to 1. This degree goes from 0 to 1. Depending on the position of an element on the horizontal axis, the value determined by the line drawn indicates a different degree of belonging. When a Universal is located in the maximum degree of belonging to that set, we can speak of that Universal being representative of that set.

That is to say, although there are Universals that share a set, their degree of belonging to it may not be the same. Although a rule present in 52% of the languages and one present in 41% of

the languages share a set and have an average Universality value, their set membership varies. The 52% rule would have a set membership value of 1. The 41% rule would not have this maximum value and, therefore, is not a prototypical representative of its group.

In studying Universals, we must remember that we carry out approximations of the phenomenon under study through large samples of languages. These can be better or worse. In the case of prototypical Universals, we can be more confident of their belonging to the corresponding set. In the case of Universals that do not exhibit the maximum degree of belonging to the set they are found, it would be interesting to carry out further checks that allow us to reinforce their belonging to the set.

On the x-axis, we can see a scale from 0 to 100. This can be modified depending on the universe. In the case of the Universals we are analyzing, this universe should be the total number of languages available in the version of Universal Dependencies analyzed. However, it should be remembered that the Universals analyzed, being implicational, change the number of languages they affect. That is, 143 languages are not analyzed each time. If the condition of the Universal is applied exclusively to languages of the basic order Verb-Subject-Object, for example, the number of languages analyzed would be much smaller. We have, therefore, decided to homogenize the scale to 100 to be able to work with percentages. We have decided to do so in order also to be able to present the percentages in case the reader is interested in them. The model does not focus on this precision of percentages but rather the opposite. However, we must comment that one of the advantages of the model is precisely that we could change the data on this horizontal axis in each Universal depending on the number of languages to which it applies and, even so, an exact comparison between Universals could be carried out, since the interesting thing is to see the point on the scale where they are located. Changing the x-axis figures is possible in other versions and would not alter the distribution of the sets at all.

16.2 Example of Application to the Universal 13

In order to explain the specific functioning of the GLU Model, we have decided to exemplify its application to the Universal 13, which we will analyze later and rescue in figure 16.2. This Universal states: “If the nominal object always precedes the verb, then verb forms subordinate to the main verb also precede it”.

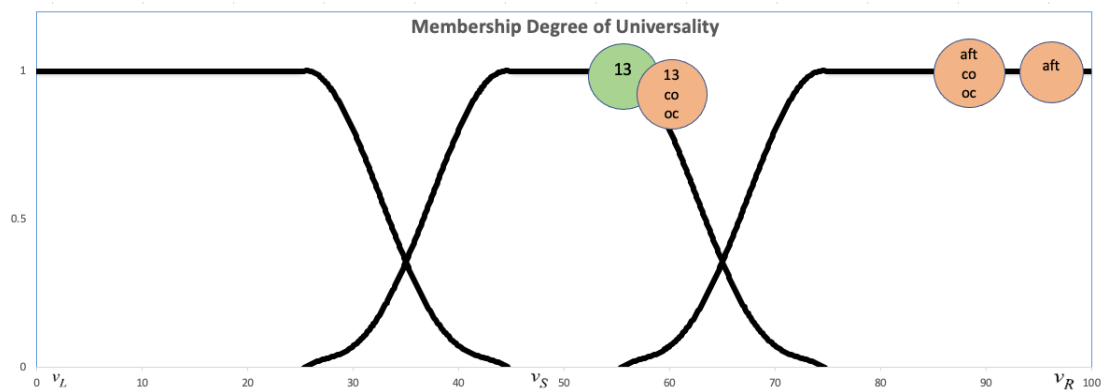


Figure 16.2: Trichotomous Expression of the GLU-M for Universal 13

The layout we chose to mark the location of the Universals on the scale is that of circles. We propose a color code to help easily identify the different natures of the claims tested. This can be modified at the discretion of the researcher. In this thesis, we use:

- **Green.** We assign this color to the proven formalizations corresponding to Greenberg's original formulation. The claims expressed with green circles are the ones that show the realization of the original Universal.
- **Blue.** Some of Greenberg's original Universals contain a complex structure, i.e., different structures. The fulfillment of the Universal depends on the individual fulfillment of these different rules. We have chosen to offer in an isolated and concrete way the individual fulfillment of each of the claims contained within a Universal with more than one formulated rule. This check makes it possible to find out whether the concrete position of an original Universal (green) is highly conditioned by the concrete location of one of its elements (blue).
- **Orange.** In this thesis, we have decided to carry out parallel checks to Greenberg's original one. These checks are based on the corresponding Universal but check other aspects that potentially can be related to the structure analyzed by Greenberg. These kinds of checks help to look at a broader picture and to obtain possibilities of extending a Universal to more structures, proposing a different Universal in case of a better location on the scale, etc.
- **Yellow.** This color is reserved for possible checks that escape the above points. This color can be considered a wild card to locate different structures that must be detailed. An example of such a case in our checking is Universal 14. In such a Universal, in addition to the checks of the original Universal and alternative proposals in categorical terms, we decided to explore a possible formulation of the Universal in terms of Typometrics. This means that we do not

work with the labels corresponding to each language type. However, we seek to observe the pure quantitative behavior and distribution of languages. This type of analysis is entirely different from the one carried out by Greenberg (and therefore also by us in comparison with Greenberg), which is why we highlight it with this special color.

We place the different circles on the line of each set instead of on the horizontal axis. This is done to facilitate the reading of the results, especially when circumferences coincide in a similar position. In such a case, the circumference is located at a lower position than the line of the parallel set. In such cases, the degree of belonging to the set must be interpreted, taking as a reference the line to which each circumference corresponds concerning the coincidence of the vertical and horizontal axes.

In the unlikely event that a rule falls right between the intersection of two sets, we would assign it to the more extreme set (i.e., low or high) for explainability reasons. To favor natural transitions and not to be highly conditioned by the sampling of languages and that the addition of one more or less language conditions the results, once a rule has been placed in a set, it is considered to belong to this set until it ceases to belong to this group completely in further checks. However, it should not be forgotten that, in these cases, a rule has a membership degree higher than 0 in two sets.

The results we have obtained after the computational analysis are as follows:

- **Greenberg's original Universal term.** That is, languages that can be considered that, if they place the nominal object before the verb, the verb forms dependent on the main verb also precede it: 57%.
- **Correlation check of the opposite direction.** That is, languages that can be considered that, if they place the nominal object after the verb, the verb forms dependent on the main verb also appear after it: 97%.
- **Testing Greenberg's original Universal in co-occurrence.** That is, analysis of cases where we can find nominal objects of a verb and that verb dependent on another head verb in the same structure. In this case, both dependents appear before the head: 60%.
- **Checking the correlation of the opposite direction in co-occurrence.** That is, analysis of cases where we can find nominal objects of a verb and that verb dependent on another head verb in the same structure. In this case, both dependents appear after the head: 88%.

Once we have these precise results, we can use them as input for our GLU Model. The output we obtain is shown in the figure 16.2. The original Greenberg Universal and the check of the same direction in cases of co-occurrence of both elements are placed in the central set. The proposal we make with the opposite direction to Greenberg and the check of that same direction in cases of co-occurrence of both elements are placed in the right-hand set. This means that in the first two cases, we are dealing with a Mid-Universal, and in the last two cases, with a High-Universal. In this case, we recommend referencing the Universal we have proposed instead of the one proposed by Greenberg. We recall that only high Universals are relevant for the Typology and classification of languages. It is also interesting to note the similar behavior of co-occurrence cases. It seems that the simultaneous occurrence does not modify the behavior of the languages, and there is no clear influence on the consistency of the direction in the same context to facilitate processing. This consistency already appears in cases of non-co-occurrence, especially in VO-directional languages.

16.3 Comparison of Traditional Labels and the GLU Model

The three fuzzy sets of the scale are low, medium, and high (from left to right). Each has some characteristics and a link to what has been observed in the literature.

The Low-Universal set shows rules that are not present in most languages. Within this set are the *rara* and *rarissima* labels. We have not proposed their detailed embedding nor exemplified their treatment in this thesis because it is not the main interest: we have devoted ourselves to just the opposite, the Universals. However, although the full correspondence and functioning of the low Universality set in real cases remains to be verified, we consider it positive that the GLU Model offers a set to make these structures available.

The Mid-Universal set shows rules found in approximately half of the languages analyzed. Within this set, we find the maximum variation and the least restriction on what languages can exhibit. Any rule that falls into this set is of no interest to Linguistic Typology, as it offers no relevant information about invariance or restriction within the variety of the world's languages.

If we establish a two-part division of this set, in the part closest to the high Universal set, we should place the statistical Universals proposed by Kirkegaard (2016). We have already commented in the literature review that this researcher proposes that any rule that exhibits more than 50% adoption in the languages analyzed is a statistical Universal. We have shown our rejection of such an approach. In the case of the GLU Model, our opposition can be seen perfectly and graphically.

A rule shared by half of the languages analyzed + 1 does not represent a clear trend in the world's languages. Moreover, the high conditioning of sampling may mean that the result in another analysis would be half of the analyzed languages - 1. This, added to the crisp conception of the proposal of Kirkegaard (2016), makes the study of Universals difficult.

The High-Universal set shows rules found in most of the languages analyzed. This set is useful for studying invariance within the variation of the world's languages. The rules in this set are equivalent to any element that would be placed in the high set in a trichotomous expression.

Therefore, this High-Universal Fuzzy Set is equivalent to what can be found in the literature under the label of statistical Universal. In this case, we find the difference that we do show the location of this set (something we have found difficult to do in the literature) with its delimitations. The set is fuzzy, so no longer belonging to this set is a matter of degree. The set eliminates the opposition between statistical and absolute Universals: absolute Universals are encompassed within the set. The distinction of exceptions is not considered to be relevant. This is a fact shared by many Typology researchers, but to our knowledge, no model has been offered that includes it. If any researcher really wants to pursue this distinction, the information is salvageable in the model: an absolute Universal is a Universal placed at the 100 point on the scale. However, as we commented, such a distinction is irrelevant to us, and we do not give relevance to using the label absolute (or exceptionless in some papers).

Through this merging of the concepts of absolute and statistical, we are, de facto, affecting many other labels that we have observed in the literature, which are euphemisms for statistical Universals. We refer to expressions such as “prediction”, “tendencies”, “typological generalizations”, “preferences”, “trends”, “probabilistic Universals”, “cross-linguistic generalizations”, “cross-linguistic patterns”, “statistical generalizations”, “linguistic tendencies”, “Universal generalizations”, “statistical patterns”, “probabilistic patterns” or “cross-linguistic tendencies”, for example. All these expressions can be synonyms for High-Universals in their broadest sense. We understand that they refer to claims that are shared by most languages, although with the possibility that some languages do not share them. In other words, we are talking about rules with a high value of Universality. In our case, we do not consider it relevant whether their high Universality value is 1, 0.99, or 0.96 since they still have a high Universality value, which is what interests Linguistic Typology.

Another label we reduce in our proposal to simplify existing terminology is “near Universal”. We consider such a distinction irrelevant, especially if it is made using crisp criteria, as done in the

literature. Our model does not include such a distinction, but it is possible to make it if a researcher wants to do so. The near Universals could be placed in a distinction within the High-Universals set, placing them in the old shift between very high and high of the 5-set models. If such a distinction is made within the High-Universals set, congruence with membership of the general High set is guaranteed, with the greater precision of indicating that it also exhibits a higher property of that same High. One could do something similar to what is proposed by Torrens-Urrutia, Novák, and Jiménez-López (2022) in their study concerning small and very small. In short, our model rules out the distinction between statistical, almost absolute, and absolute Universals. At the same time, we also contribute to discard labels such as “nearly exceptionless” or “significant statistical”, which we understand to be synonymous with “almost absolute Universals”.

Other tags seen as “almost trivial” are also no longer necessary in our present model. In this case, we can understand them as rules shared by most languages but with a high number of exceptions. This definition may correspond to High-Universals that do not have a maximum degree of membership degree to their set. Therefore, we are dealing with Universals which, we must not forget, still exhibit a clearly high degree of Universality (shared by most languages) but whose high number of exceptions makes them fall somewhere in the High-Universal fuzzy set and near the Mid-Universal fuzzy set.

16.4 Benefits of the GLU Model

The model we have presented contains some positive points that make it interesting for addressing the objectives we are concerned with. We can speak, in general, of the following aspects:

1. GLU allows the integration between statistical and absolute Universals.

The literature review shows how Universals in generative grammar are always absolute or cannot be Universals. In the case of Linguistic Typology, multiple authors have spoken of the irrelevance of the distinction between absolute and statistical Universals in this discipline. The exceptions found in the realizations of some languages may be welcome. However, we cannot be sure it is such in the case of having an absolute Universal.

Therefore, the model allows us a gradual classification of the Universals. The scale on which they are placed reflects the greater closeness between a Universal with an exception or with none, practically in the same position. This naturalization is related to what Cook (2002) points out about how the small differences in the assignment of a real number are artefactual.

He comments that a subject with a value of 1 in tall and a subject with a value of 0.97 in tall, for example, can be considered tall. For him, the assignment of one value or the other is small and, therefore, not representative of anything. In other words, the fact that a claim is not true in 2 or 3 languages, for example, does not invalidate its status as a real trend in the world's languages, worthy of consideration from the point of view of Linguistic Typology. It is more natural, in our opinion, this consideration, which deals with the injustice of having a label for a claim with no exceptions found (for the moment) and another label for the range of languages from 1 exception to a number that has not been determined in the bibliography either.

2. GLU allows the reduction of the terminology used.

If we carry out the previous point, the distinction between absolute and statistical Universals no longer makes sense. Neither do the multiple labels detected in the literature review, such as “statistical tendency”. It may be beneficial to stop using these multiple labels referring to the same problem to gain clarity in the description of Universals.

3. GLU allows for greater consistency with label changes.

If the distinction between statistical and absolute Universals is maintained, as mentioned above, assigning one label or the other depends on detecting exceptions. Therefore, a Universal considered absolute may become statistical with another sample of languages. Also, conversely, a language that has been presented as an exception may not be such under another interpretation of its linguistic behavior in the future. In such a case, the Universal would change from statistical to absolute. Since typologists do not attach importance to the presence of some exceptions, this change of labels highly dependent on other elements is not desirable.

4. GLU allows a gradual classification of Universals.

It is not easy to find clear proposals that distinguish between what is Universal and what is not, especially without sharp boundaries. In the model presented, in addition to clearly presenting a set corresponding to the High-Universals, which are the Universals of interest, this is done gradually. This allows for less dependence on the concrete result obtained by the data and more adaptation as more data becomes available.

This feature is especially interesting in computational analyses such as the one we have carried out. The sample of languages in Universal Dependencies is imperfect and could show a greater variety and representativeness in some areas. Therefore, we understand that the quantitative

result of the Universals may vary slightly when adding more desired and necessary languages. However, given the flexibility of the proposed model, a set change would only occur if there was a radical change in the data obtained in another sampling, which seems unlikely.

5. GLU allows a quick view of the results.

In its interface, we consider that the model used is quite visual and allows easy reading of the group belonging to the Universals in case this is the only thing that interests the reader. It also allows us to place multiple Universals on the same scale or compare them with a different extension on their horizontal axis. On the other hand, if you want more data, such as the prototypicality value within its set (van der Auwera & Gast, 2010), it is also possible to have this information.

6. GLU is flexible to each researcher's needs.

As mentioned above, our main objective is the identification of claims that represent invariance within language variation. We understand this as equivalent to High-Universals. However, other researchers may be interested in analyzing claims corresponding to Low-Universals or researchers who consider it necessary to distinguish subsets within High-Universals. In previous sections, we have shown how this is possible and how the model can be adapted to the research needs, just as we believe that what we offer is appropriate for the research we carry out.

7. GLU allows the classification of linguistic types.

The model we use has many other applications in linguistics, specifically in Linguistic Typology. Another of the options that we see as interesting (although it was not the object of study of this thesis) is the classification of linguistic types through this model. This would allow us to represent graphically, in an obvious and evident way, the behavior of languages in a continuum and, at the same time, to offer a grouping in sets that is also useful, something that would benefit what Guzmán Naranjo and Becker (2018) or Grandi (2020), who point out that pure types do not exist, claim. For example, one could take the figure 16.2 as the representative of any type; in the two extremes, one can place the two directions in word order of that dependency, while the central set would be for the representation of the non-dominant order.

16.5 On Possible Objections to the GLU Model

The trichotomous expression we have presented focuses on representing the object of study in three fuzzy sets. The transition from one to another is not abrupt but gradual. It is not the exact result that matters but the place where the Universal is located in the set.

In contrast, when we analyze the Greenberg Universals, precise quantitative data about the result of each Universal in the set of language corpora analyzed can be seen. Although the exact quantitative score is not interesting in our adopted model, it is fundamental and necessary for the model's functioning. In other words, in a fuzzification process, we can have a crisp or fuzzy value as input, but the important thing is that the output is always a fuzzy set. In other words, we are interested exclusively in the output we are looking for, and we use precise and concrete data as input for the model.

Since we need such data, we have thought it convenient to offer the result obtained, which may be interesting information for the reader who is interested in precise quantitative results. In addition, this way, we allow better traceability of the data.

This apparent contradiction has been discussed at length by Cook (2002:225): “One of the main reasons for providing formal semantics for languages is that the mathematical precision afforded by such semantics allows us to study and manipulate the formalization much more easily than if we were to study the relevant natural languages directly”.

Cook (2002:225) talks about “mathematical precision” as this exact quantitative result. He comments that in order to facilitate the manipulation of the model, we often need to include these types of elements:

With this distinction in place, the criticisms of Sainsbury and Tye are efficiently dealt with. The precision of the semantics is artefactual and does not represent any real precision in vague discourse. Although this solution to this problem is independent of any particular semantics, a detailed account of how we would distinguish between representor and artefact within Dorothy Edgington's degree-theoretic semantics.

In the above quote, we can identify the use of “artifact” as fundamental. This should be understood as all the elements inherent to the model and that it needs to function, which are not a “representor” of the model. That is to say, it should not be understood that the mathematical precision used represents the nature of the vague element we are trying to model. A very clear example is that, by adding or deleting some language from the set, the result of the Universal could be a few tenths different. Therefore, its mathematical precision would change. On the other

hand, the Universal would still be part of the same set in which it was previously, which is what we are interested in knowing.

This makes us see that it is difficult to assert the nature of a Universal in terms of mathematical precision but not in terms of the set where we seek to place it. In other words, it seems that it is not possible to say that a rule is 89% Universal since this data is highly conditioned by the sample used. On the other hand, it may make sense to assert (from the 89% obtained) that this rule is a High-Universal. Similar results will be obtained in other types of samples, but the group will not change. Therefore, there is less sampling conditioning.

A similar criticism to that of mathematical precision could be made about the cutting points existing in the model (Sainsbury, 1990; Tye, 1994). Although there is a gradual transition from one set to another, there is, of course, a point at which there is a cut-off point at which a set ceases to exist on the scale. This apparent contradiction between the Sharp boundaries existing in the model and with the gradient modeling of a phenomenon is not such according to Cook (2002:230):

At least some parts of any model are intended to be, in some sense, important, representing (in a perhaps simplified way) real aspects of the phenomenon being modeled. However, other parts of the model might not be intended to match up with anything real. In other words, although we require that some parts of the model match up with some aspects of the phenomenon, not all have to do so.

Therefore, Cook (2002) understands this case as another example of the differentiation between elements of the model that are representative of aspects of the modeled object and elements without such correspondence. Elements necessary for the model's functioning but whose intent or use does not go beyond this is what has been defined above as artifact (Keefe, 2012). In other words, a mathematically accurate result such as 89.21% is an artifact needed by the model to achieve placement in the High-Universal set, which is a representor in that it gives us the information we need from the modeled element.

Therefore, it should not be surprising that any many-value model naturally presents several super-precise borders. The goal is not to provide an "accurate description of the semantics of vague languages" (Tye, 1994), but to model a vague linguistic aspect. This has also been advocated by Shapiro (2006:53):

The objection is misguided against the logic-as-model approach to logic. From that perspective, the precision of the continuum to model the fuzzy slide from truth to falsity as we go through a sorites series. There is nothing untoward about using a precise structure to model a vague one (see also Gaifman 2005). The model theory developed here has this feature as well.

By modeling through logic, we mean showing how a vague linguistic aspect behaves. The formalism is not a description of what is happening but a fruitful way of representing the phenomenon (Cook, 2002:234). Therefore, not all aspects inherent in the model must necessarily correspond to aspects of the modeled phenomenon.

Part V

A Computational Approach to Greenberg's Universals

UNIVERSITAT ROVIRA I VIRGILI

GLU - A GRADIENT MODEL FOR LANGUAGE UNIVERSALS. A COMPUTATIONAL REVISITATION OF GREENBERG UNIVERSALS

Antoni Brosa Rodríguez

17

Formal Prerequisites

Contents

17.1	Universal Dependencies Tags	228
17.2	Surface Universal Dependencies Syntactic Relations	241
17.3	Grew-Match Formalisation Syntax	243
17.4	Data Management Formalisations	244

Below are all the symbols necessary to understand the formalism we propose. We must warn that not all the symbols to be presented have been used, as some of them are specific to some language groups or are not necessary for this thesis. Nevertheless, we intend to show the complete functioning of the formalization proposal, as it is a valuable output and intended to be a starting point for future work by the researcher responsible for this doctoral thesis (and perhaps for other linguists) in other works.

The labels for the parts of speech and the various morphological features are those proposed in Universal Dependencies. These are the ones used in Grew-Match and enjoy the widest acceptance in the community. All examples are extracted from Universal Dependencies (or Surface Universal Dependencies), where you can go for further information. For the reasons commented throughout the work, the labels that will be used for the dependency relations in syntax are based on Surface Universal Dependencies (SUD).

The rest of the elements necessary for the formalization (the syntax of the formalism) are those required by Grew-Match for its operation. All these elements will be presented, and, in addition, the specific tags that can be used to obtain extra functionalities in this tool will also be shown.

We also summarize the different elements included in the formalizations we created in the Data Management section.

17.1 Universal Dependencies Tags

The Universal Dependencies annotation scheme includes different tags for different morphosyntactic aspects. The tags used to annotate the parts of speech, based on Google’s proposal (Petrov et al., 2012), are among the most widely recognized and used. These are often based on the classical conventions that abbreviate the different parts of speech. The distinction between PROP_N and NOUN is particularly noteworthy, i.e., between proper and common nouns. A distinction is also made between coordinating conjunctions (CCONJ) and subordinating conjunctions (SCONJ). It is important to add that there is a specific tag to the symbols (SYM) and a tag called X that allows the addition of any existing information that does not fall within the classic categorization, such as the presence of noises or unintelligible words. All this is collected in the table 17.1.

Part of Speech Tag	Part of Speech Value	Example
ADJ	Adjective	red shoes
ADV	Adverb	arguably wrong
INTJ	Interjection	ouch
NOUN	(Common) Noun	girl
PROP _N	Proper Noun	London
VERB	Verb	ate
ADP	Adposition	in London
AUX	Auxiliary	was done
CCONJ	Coordinating Conjunction	or
DET	Determiner	this car
NUM	Numeral	6
PART	Particle	John’s car
PRON	Pronoun	What do you think?
SCONJ	Subordinating Conjunction	if
PUNCT	Punctuation	,
SYM	Symbol	\$
X	Other	And then he just xfgh pdl jklw

Table 17.1: Universal Part of Speech

On the other hand, we have all the morphological information that can be annotated thanks to the Universal Dependencies annotation scheme. Again, the aim is to be as universal as possible. Some categories are included that are not genuinely universal and apply to some specific languages, such as NounClass. However, it is necessary to have these tags to capture all the information in those languages, another of the objectives of UD. Some tags allow you to mark words with an alternative status, such as abbreviations (Abbr), foreign words (Foreign), or misspellings (Typo). In short, all the information that UD theorizes to be annotated as features of words can be found in the table 17.2.

Feature Tag	Feature Value	Feature Tag	Feature Value
PronType	Pronominal Type	Definite	Definiteness or state
NumType	Numeral Type	Degree	Degree
Poss	Possessive	VerbForm	Form of verb or deverbative
Reflex	Reflexive	Mood	Mood
Foreign	Foreign Word	Tense	Tense
Abbr	Abbreviation	Aspect	Aspect
Typo	Typo or misspelled word	Voice	Voice
Gender	Gender	Evident	Evidentiality
Animacy	Animacy	Polarity	Polarity
NounClass	Noun Class	Person	Person
Number	Number	Polite	Politeness
Case	Case	Clusivity	Clusivity

Table 17.2: Features

If we develop the features in the order they are presented in UD, we have the category of pronominal type. This does not correspond precisely to a pronoun, although it encompasses them. It should be noted that UD considers the functioning of the comparative objects shown above. This means that classical categories are used with much more flexible and extensive content, which includes similar behavior in different languages. Thus, we have the PronType “Art”, for example. It should be understood that there is no possible confusion because of the grammatical category of “the”. However, it has the value of PronType Art and has the part of speech of DET (determiner), not of PRON (pronoun). In summary, the different values that exist for PronType are listed in the table 17.3.

The numerals have their own category (NumType), as there are different types. In some cases, this detail level is unprecedented, such as the possibility to distinguish between ranges and sets. The precision in the numeral type can potentially allow further refinement of the linguistic

PronType Tag	PronType Value	Example
Art	Article	the
Dem	Demonstrative pronoun, determiner, numeral or adverb	here
Emp	Emphatic determiner	He himself did it.
Exc	Exclamative determiner	What a surprise!
Ind	Indefinite pronoun, determiner, numeral or adverb	Few people
Int	Interrogative pronoun, determiner, numeral or adverb	Who are you?
Neg	Negative pronoun, determiner or adverb	No
Prs	Personal or possessive personal pronoun or determiner	She
Rcp	Reciprocal pronoun	einander (German)
Rel	Relative pronoun, determiner, numeral or adverb	jenž (Czech)
Tot	Total (collective) pronoun, determiner or adverb	všechno (Czech)

Table 17.3: Pronominal Type

Universals, which have remained merely at the general numeral level. The classification proposed by UD is based on seven numeral possibilities listed in the table 17.4.

NumType Tag	NumType Value	Examples
Card	Cardinal number or corresponding word	two
Dist	Distributive numeral	Három-három (Hungarian)
Frac	Fraction	three-quarters
Mult	Multiplicative numeral or corresponding word	double
Ord	Ordinal number or corresponding word	second
Range	Range of values	two-five (‘two to five’)
Sets	Number of sets of things: collective numeral	two/three pairs of shoes

Table 17.4: Numeral Type

On the other hand, the possessive feature does not contain different values or nuances. We only have the possibility of having a feature indicating possession or not. Therefore, there is only one possibility. This refers to having a possessive value, which has been named “Yes”. Its opposite is the absence of value; therefore, it does not need a label, as the possessive feature will not be activated, as shown in the table 17.5.

Poss Tag	Poss Value	Example
Yes	The only value is the presence of possessive	Her house

Table 17.5: Possessive

The same is true of whether a word has the characteristic of reflexivity. There is no more nuance than the fact of being so or not. In this case, only one tag for the “Reflex” value, “Yes”. If the word has no reflexive value, the “Reflex” feature is not activated, as shown in the table 17.6.

Reflex Tag	Reflex Value	Example
Yes	The only value is the presence of reflexive	Si (Italian)

Table 17.6: Reflexive

The foreign word tag also supports only the possibility of existing or not without further detail. This is because it is not very productive to detail the origin of this foreign word, which would involve thousands of tags. Although it may be of little interest, from a general perspective, to tag such constructions, their use in specific studies can be very interesting in specific investigations. Tagging these constructions would allow, for example, to check unequivocally and quantitatively language trends, such as the greater facility of Italian for the adoption of foreign words in comparison with other Romance languages such as Catalan. This fact is summarised in the table 17.7.

Foreign Tag	Foreign Value	Example
Yes	The only value is the presence of a foreign word	He said I could dra åt helvete!

Table 17.7: Foreign Word

Universal Dependencies have also been chosen to show whether the form of a word is full or abbreviated. If it is abbreviated, it will display a “Yes” as the value of the “Abbr” feature, equivalent to an abbreviation. Again, this information may seem unproductive for the study being carried out in this thesis. However, this tag makes possible detailed and quantitative analyses of the behavior of languages about this aspect, which has been little treated in disciplines such as Typology. Therefore, the table 17.8 shows how this is summarised.

Abbr Tag	Abbr Value	Example
Yes	The only value is the presence of an abbreviation	UK

Table 17.8: Abbreviation

Another example of a tag with a single value is the “Typo” tag. If the text has some kind of obvious typo, it can be signaled by activating this tag’s single value (Yes). Therefore, we summarise the possibilities in the small table 17.9.

Typo Tag	Typo Value	Example
Yes	The only value is the presence of a typo	Barak Obama

Table 17.9: Typo or Misspelled Word

Concerning the different gender values, we do find more variety. In addition to the masculine and feminine genders, the neuter and common genders are considered relevant tags for a comprehensive representation of the languages. These are listed in the table 17.10.

Gender Tag	Gender Value	Example
Com	Common gender value	väg (Swedish)
Fem	Feminine gender value	Burg (German)
Masc	Masculine gender value	hrad (Czech)
Neut	Neuter gender value	barn (Swedish)

Table 17.10: Gender

Different languages around the world show different ways of marking animacy feature values. From the annotation scheme proposed by UD, it is understood that four main categories result from the affirmation or negation of the animacy value and the humanity value of the element in question. Therefore, the existing tags are those in the table 17.11.

Animacy Tag	Animacy Value	Example
Anim	Animate value	malí psi (Czech)
Hum	Human value	mali chłopcy (Polish)
Inan	Inanimate value	małe domy (Polish)
Nhum	Non-human value	małe psy (Polish)

Table 17.11: Animacy

UD has also decided to distinguish the Noun Class feature. This responds to a special marking of the type of noun, with characteristics that are different from one another, being very specific. In our opinion, these classifications are not very universal, as they are not widely applicable to the world's languages. A clear proof of this is how these tags are labeled, with the name of the language or family that needs them (Bantu or Wolof). Therefore, we understand that it is not an interesting universal tag for studies such as the one we propose. However, we understand it is an accurate, sensible, and practical tag for the languages concerned. It may be helpful for you to provide an answer on how to annotate such specific elements. In any case, we provide their nomenclature to be exhaustive and represent all tags available in UD, whether they are used in this thesis or not. The proportion of the complete annotation scheme allows us to extend the methodology used here to other types of typological studies that may be of interest even though they have not been foreseen in this study. The table that collects the different tags referring to

the NounClass values are those of the table 17.12.

NounClass Tag	NounClass Value	Example
Bantu1	singular, persons	mtoto (Swahili)
Bantu2	plural, persons	watoto (Swahili)
Bantu3	singular, plants, thin objects	mti (Swahili)
Bantu4	plural, plants, thin objects	miti (Swahili)
Bantu5	singular, fruits, round objects, paired things	jiwe (Swahili)
Bantu6	plural, fruits, round objects, paired things	mawe (Swahili)
Bantu7	singular, things, diminutives	kitabu (Swahili)
Bantu8	plural, things, diminutives	vitabu (Swahili)
Bantu9	singular, animals, things	ndege (Swahili)
Bantu10	plural, animals, things	ndege (Swahili)
Bantu11	long thin objects, natural phenomena, abstracts	utoto (Swahili)
Bantu12	singular, small things, diminutives	akabwa (Swahili)
Bantu13	plural or mass, small amount of mass	otuzzi (Swahili)
Bantu14	plural, diminutives	obubwa (Swahili)
Bantu15	verbal nouns, infinitives	kusoma (Swahili)
Bantu16	definite location, close to something	pahali (Swahili)
Bantu17	indefinite location, direction, movement	kule (Swahili)
Bantu18	definite location, inside something	mule (Swahili)
Bantu19	little bit of, pejorative plural	hyundu (Hunde)
Bantu20	singular, augmentatives	ogusajja (Ganda)
Bantu21	singular, augmentatives, derogatives	dinga (Venda)
Bantu22	plural, augmentatives	agasajja (Ganda)
Bantu23	location with place names	elugala (Ganda)
Wol1	1/k (singular human)	ki (Wolof)
Wol2	2/ñ (plural human)	ñi (Wolof)
Wol3	3/g (singular)	gi (Wolof)
Wol4	4/j (singular)	ji (Wolof)
Wol5	5/b (singular)	bi (Wolof)
Wol6	6/m (singular)	mi (Wolof)
Wol7	7/l (singular)	li (Wolof)
Wol8	8/y (plural non-human)	yii (Wolof)
Wol9	9/s (singular)	si (Wolof)
Wol10	10/w (singular)	wa (Wolof)
Wol11	11/f (location)	fi (Wolof)
Wol12	12/n (manner)	ni (Wolof)

Table 17.12: Noun Class

The next feature from the alphabetical order is the number. In this case, there are 11 different possibilities, some more widespread in languages than others. All of them allow us to capture well the number generation systems of the world's languages, even those with a high specificity. The different values proposed to annotate any language in terms of number are those in the table 17.13.

Number Tag	Number Value	Example
Coll	Collective, mass, singulare tantum	lidstvo (Czech)
Count	Count plural	tri stola (Bulgarian)
Dual	Dual number	glasova (Slovenian)
Grpa	Greater paucal number	gimhat (Sursurunga)
Grpl	Greater plural number	geseeli (Fulfulde)
Inv	Inverse number	ę'gò sạ: dò è khópdò (Kiowa)
Pauc	Paucal number	karlipatu (Warlpiri)
Plur	Plural number	cars
Ptan	Plurale tantum	scissors
Sing	Singular number	car
Tri	Trial number	skobiser (Biak)

Table 17.13: Number

Cases are another feature that can be perfectly marked in UD. Cases can be multiple, as there are languages with very rich systems. Others exhibit much more specific cases, even if they are not so rich in this aspect, which must also be represented. This causes the number of cases to be quite broad, in order to represent and collect any type of case from the world's languages and fit it into one of the categories proposed by the UD annotation scheme. In any case, many of these tags, being quite precise for certain linguistic families, will not be very productive in the field of Universals. However, their typological and linguistic interest is undeniable. The summary of the available tags, their value, and an example can be found in the table 17.14.

Case Tag	Case Value	Example
Abs	Absolutive	Maria lotan dago (Basque)
Acc	Accusative / oblique	He loves her
Erg	Ergative	Juanek Maria maite du (Basque)
Nom	Nominative / direct	She sleeps
Abe	Abessive / caritive / privative	rahatta (Finnish)
Ben	Benefactive / destinative	multilarentzat (Basque)
Cau	Causative / motivative / purposive	jokaeragatik (Basque)
Cmp	Comparative	phulāpēksā(Marathi)
Cns	Considerative	miyiwana (Warlpiri)
Com	Comitative / associative	koeraga (Estonian)
Dat	Dative	Ich gebe meinem Bruder (German)
Dis	Distributive	fejenként (Hungarian)
Equ	Equative	bence (Turkish)
Gen	Genitive	Praha je hlavní město České republiky (Czech)
Ins	Instrumental / instructive	psát perem (Czech)
Par	Partitive	kolme taloa (Finnish)
Tem	Temporal	héktor (Hungarian)
Tra	Translative / factive	kella kuueks (Estonian)
Voc	Vocative	Co myslis, Filipe? (Czech)
Abl	Ablative / adelative	rannalta (Finnish)
Add	Additive	riiki (Estonian)
Ade	Adessive	laual (Estonian)
All	Allative / adlative	pénztárhoz (Hungarian)
Del	Delative / superlative	asztalról (Hungarian)
Ela	Elative / inelative	talosta (Finnish)
Ess	Essive / prolative	lapsena (Estonian)
Ill	Illative / inlative	taloon (Finnish)
Ine	Inessive	talossa (Finnish)
Lat	Lative / directional allative	etzerantz (Basque)
Loc	Locative	Mluvili jsme tam o morfologii (Czech)
Per	Perlative	yurutuwana (Warlpiri)
Sbe	Subelative	mixinikaj (Lezgian)
Sbl	Sublative	bokorra (Hungarian)
Spl	Superlative	vičeldi (Lezgian)
Sub	Subessive	tarcin serindik (Lezgian)
Sup	Superessive	asztalon (Hungarian)
Ter	Terminative / terminal allative	kella kuueni (Estonian)

Table 17.14: Case

Being definite or not is also an interesting feature in linguistics. However, in this case, it is not possible to count only with an active value of “Yes”. The possibility of being indefinite should also be explicitly mentioned, as well as the possibility of picking up different nuances that occur in languages such as Arabic. The existing options for UD are those in the table 17.15.

Definite Tag	Definite Value	Example
Com	Complex	al-muxtalifatu (Arabic)
Cons	Construct state / reduced definiteness	hulwatu (Arabic)
Def	Definite	the dog
Ind	Indefinite	a dog
Spec	Specific indefinite	c’ą wą ‘ag.li (Lakhota)

Table 17.15: Definiteness or State

The possible degree features are distributed in seven different values. They are mainly applied to adjectives. Some of the tags we will show are also applicable to nouns. This degree of specificity allows a more detailed typological knowledge than usually exists. The summary of all the possibilities, with the tag used in UD and a representative example of each type, appears in the table 17.16.

Degree Tag	Degree Value	Example
Abs	Absolute superlative	guapísimo (Spanish)
Aug	Augmentative	chlapak (Czech)
Cmp	Comparative, second degree	the man is younger than me
Dim	Diminutive	appeltje (Dutch)
Equ	Equative	pikkune (Estonian)
Pos	Positive, first degree	young man
Sup	Superlative, third degree	this is the youngest man in our team

Table 17.16: Degree

We also have different forms or possibilities in the case of verbs (or deverbative categories). These usually correspond to the traditional nomenclatures, as seen in the table 17.17. Their application extends to words that may have verbal behavior without being purely verbal in some languages.

VerbForm Tag	VerbForm Value	Example
Conv	Converb, transgressive, adverbial participle, verbal adverb	zírali na mne, pevne svírajíce sve zbrane (Czech)
Fin	Finite verb	he does
Gdv	Gerundive	puer laudandus est (Latin)
Ger	Gerund	I look forward to seeing you
Inf	Infinitive	Ich muss gehen (German)
Part	Participle, verbal adjective	he could have been prepared if he had forseen it
Sup	Supine	Jag har ätit maten (Swedish)
Vnoun	Verbal noun, masdar	delani (Czech)

Table 17.17: Form of Verb or Deverbative

On the other hand, we have the verbal mode. This is concretized in 14 categories in the UD annotation scheme, detailed in the table 17.18. Some of these are specific to some languages, such as those we provide with the example of Turkish. Therefore, in these cases, the annotation is relevant to adequately and accurately describe these types of languages.

Mood Tag	Mood Value	Example
Adm	Admirative	Ti fliske shqip! (Albanian)
Cnd	Conditional	eve gidiyorsa (Turkish)
Des	Desiderative	tabetai (Japanese)
Imp	Imperative	eve git (Turkish)
Ind	Indicative or realis	eve gitti (Turkish)
Int	Interrogative	Nach in aghaidh easa ata se ag snamh? (Gaelic)
Irr	Irrealis	Xaq ta ne kimb'e iwuk' (Quiche)
Jus	Jussive / injunctive	maivam vocah (Sanskrit)
Nec	Necessitative	eve gitmeli (Turkish)
Opt	Optative	eve gidelim (Turkish)
Pot	Potential	eve gidebilir (Turkish)
Prp	Purposive	Antha yukarnda puntayi manilhiku (Arabana)
Qot	Quotative	Sa ei tulevat (Estonian)
Sub	Subjunctive / conjunctive	Je veux que tu le fasses (French)

Table 17.18: Mood

Another typically verbal feature is the tense. The UD proposal in its annotation scheme is quite contained. Since different languages can show quite a lot of precision, in some cases,

the equivalent values are merged under the same tag. This classification makes it possible to characterize languages with a comprehensive register of tenses as those in which, for example, there is only a distinction between past and non-past tenses. All these possibilities are represented in the table 17.19.

Tense Tag	Tense Value	Example
Fut	Future tense	irá a casa (Spanish)
Imp	Imperfect	tja ostavase, kadeto toj i da otidese (Bulgarian)
Past	Past tense / preterite / aorist	he went home
Pqp	Pluperfect	afirmou que os sequestradores já ligaram (Portuguese)
Pres	Present / non-past tense / aorist (unmarked)	he goes home

Table 17.19: Tense

In the same sense, we find the development of aspect. Likewise, there are languages with very different behavior. If necessary, the UD proposal allows the application of the specific value of the aspect of the verbs. The tags they have decided to use, their value, and an example are shown in the table 17.20.

Aspect Tag	Aspect Value	Example
Hab	Habitual aspect	he attends classes of Japanese
Imp	Imperfect aspect	pekl chleba (Czech)
Iter	Iterative / frequentative aspect	utoget (Hungarian)
Perf	Perfect aspect	upekl chleba (Czech)
Prog	Progressive aspect	eve gidiyor (Turkish)
Prosp	Prospective aspect	Liburua irakurriko behar du (Basque)

Table 17.20: Aspect

Another parameter related to the verbal category is the type of voice used. In this case, there is also a wide variety of behaviors in the languages of the world, ranging from a more significant distinction to a lesser one. In order to adequately and representatively tag any possible language, UD proposes a list of 10 tags that we collect in the table 17.21.

Voice Tag	Voice Value	Example
Act	Active or actor-focus voice	Naglilinis siya ng bahay (Tagalog)
Antip	Antipassive voice	Wagu:da gudaGanda wawa:dinu (Yidiny)
Bfoc	Beneficiary-focus voice	Ipagaalis sa sako ang bata (Tagalog)
Cau	Causative voice	mosat (Hungarian)
Dir	Direct voice	Niwicihananak (Cree)
Inv	Inverse voice	Niwicihikonanak (Cree)
Lfoc	Location-focus voice	Aalisan ang sako para sa bata (Tagalog)
Mid	Middle voice	Luomai ton hippon (Greek)
Pass	Passive voice	Nililinis niya ang bahay (Tagalog)
Rcp	Reciprocal voice	Filiz ve Baris opustüler (Turkish)

Table 17.21: Voice

The annotation scheme also includes the possibility of displaying two different evidentiality values. This category is rare in the world’s languages and shows whether the expressed information is known to the sender in a direct (first hand) or indirect (Nonfirst hand) way. However, thanks to this feature, this distinction is represented in detail in the languages where it is necessary, for example, in Turkish, as shown in the table 17.22.

Evident Tag	Evident Value	Example
Fh	First hand information	geldi (Turkish)
Nfh	Non first hand information	gelmis (Turkish)

Table 17.22: Evidentiality

The polarity feature allows the addition of the value of positive or negative to the different words of the different languages that need it. This can be used in any word, whether exclusive to mark polarity or included in another word by different mechanisms. This simple distinction appears in the tab 17.23.

Polarity Tag	Polarity Value	Example
Neg	Negative polarity	no (Spanish)
Pos	Positive, affirmative polarity	sí (Spanish)

Table 17.23: Polarity

About the person to whom the words can refer, we have 5 different possibilities in UD. In addition to the 3 most frequent persons, we add a person 0 and a person 4. The person 0 label,

for UD, applies to constructions with a special mark of impersonality. The person 4 label, for UD, reflects a special third-person mark. This distinguishes a third person from a different third person in the same sentence. The example shown in table 17.24 shows an example of ambiguity between persons 0 and 4, which is quite common in languages (as it is easily disambiguated by the presence of another third person or not).

Person Tag	Person Value	Example
0	zero person	caku (Keres / Keresan)
1	first person	we
2	second person	you
3	third person	she
4	fourth person	caku (Keres / Keresan)

Table 17.24: Person

In many languages, there is a distinction about the degree of politeness of the words (verbs or pronouns, for example). There are different possibilities, depending on the language, as some of them, such as Japanese, are peculiar examples. Table 17.25 shows the different tags that can mark politeness, the value to which they correspond, and an example.

Polite Tag	Polite Value	Example
Elev	Referent elevating feature	nasaru (Japanese)
Form	Formal register feature	usted (Spanish)
Humb	Speaker humbling feature	itasu (Japanese)
Infm	Informal register feature	ty (Polish)

Table 17.25: Politeness

The last feature we highlight is called clusivity. This refers to the distinction made by some languages of the listener within the grammatical person. The tag “Ex” refers to the value of “Exclusive”, which is equivalent to the use of “we” as a sum of “I” + “they”. The tag “In” refers to the value of “Inclusive”, which is equivalent to the use of “we” as a sum of “I” + “you”. The summary of this is reflected in the table 17.26:

Clusivity Tag	Clusivity Value	Example
Ex	Excluding the hearer	kami (Indonesian)
In	Including the hearer	kita (Indonesian)

Table 17.26: Clusivity

17.2 Surface Universal Dependencies Syntactic Relations

In addition to the UD annotation scheme, which we use for the different parts of speech tags and features we have shown, we must reflect the Surface Universal Dependencies annotation scheme tags used to mark the syntactic dependency relations. To do so, we have broken them down into different tables that make them easier to read and compare with the classical dependencies. Since the SUD labeling scheme is relatively tidy and hierarchical, we first provide the general and proper SUD annotation scheme in the table 17.27. In this case, it is divided into only four general dependency possibilities. On the one hand, we find “subj” (subject) dependency. On the other hand, we find the “udep” dependency, which is used in cases where it is not clear whether it is a complement or modifier, since it is a more general dependency. There is also the “comp” tag, which is used for any type of argument. The “mod” tag is used for any type of element whose function is only to modify.

SUD Tag	Value	Example
subj	Subject	This shall be applauded
udep	Complements without distinguishing arguments and modifiers	An act of creation
comp	Arguments of verbs, nouns, adjectives, adverbs, auxiliaries, adpositions and conjunctions	Il s’en sort bien (French)
mod	Modifiers of verbs, nouns, adjectives, adverbs, auxiliaries, adpositions and conjunctions	a country with so many different languages

Table 17.27: Own Dependencies

Once it has been pointed out that some element is a dependency of type “comp”, that is to say, argumental or necessary, it can be specified what type of complement it is. To do this, the label “comp” is rescued, a colon “:” is added, and the type of comp is added. We find five types: those related to auxiliaries, those that occur in cleft sentences, those of direct objects, and those of oblique or predicative complements. All this is summarized and exemplified in the table 17.28.

Comp Tag	Comp Value	Example
comp:aux	Argument of auxiliaries	Do you remember ?
comp:cleft	Complement of cleft sentences	It was from John that she heard the news
comp:obj	Direct object complements	J’adore ce bar (French)
comp:obl	Oblique arguments	He gives you a hooded look
comp:pred	Predicative arguments of verbs	Norton had become completely disgruntled

Table 17.28: Complement Specific Dependencies

On the other hand, we have dependencies that have not been modified from the classic UD annotation scheme. We are talking about 15 more specific or less frequent tags, which are necessary for an efficient and complete annotation of corpora, such as the “vocative” (vocative) or “appos” (apposition) tag. All of them are mentioned and exemplified in the table 17.29.

SUD Tag	SUD Value	Example
vocative	Vocatives	Guys , take it easy!
compound	Compounding	ice cream
dislocated	Dislocated elements	to jani ton kserume poli kala (Greek)
discourse	Discourse elements	Iguazu is in Argentina :)
appos	Appositional modifier	Sam, my brother , arrived
det	Determiner	The man is here
clf	Classifier	san ge xuesheng (Mandarin)
conj	Conjunct	Bill is big and honest
cc	Coordinating conjunction	And then we left
flat	Flat multiword expression	Carl XVI
parataxis	Parataxis	Calafia has great fries (they are to die for!)
orphan	Orphan	Marie won gold and Peter bronze
goeswith	Goes with	For get that!
reparandum	Overridden disfluency	Go to the right- to the left
punct	Punctuation	Go home!

Table 17.29: Shared Dependencies with Universal Dependencies

Finally, we collect in the 17.30 table a further level of addition and detail in the annotation. This is done by using “@” after the previous tags and allows for greater awareness of specific phenomena, such as marking agents “agent” or signaling that it is a light verb construction “lvc”. However, practically none of the annotators of the corpora present in UD have decided to reach such a level of precision. Therefore, since their use is not currently productive, no use is made of these tags in this thesis, although they will be detailed for future research.

SUD deep tag	SUD deep Value	Example
@agent	Agent	Food was cooked by many hands
@caus	Causative auxiliaries	Il fait accélérer ses troupes (French)
@expl	Expletives	It rains a lot in England
@lvc	Light Verbs	Take part in the discussion
@pass	Passive auxiliaries	They were killed
@relcl	Relative clause	This is money you’ve earned
@tense	Argument of tense auxiliaries	This is money you’ve earned
@x	Verbs without explicitly marked subject nor semantically	I am proud to be a linguist

Table 17.30: SUD Deep Features

17.3 Grew-Match Formalisation Syntax

Once we have shown the different tags referring to all the information that can be formalized in UD, we must show the syntax we will use for these formalizations. In this case, the most reasonable thing to do is to use Grew-Match’s own formalization proposal, the tool we will use to check the data in the different corpora. This syntax will allow us to carry out the queries without any additional conversion and is perfectly suitable for any formalism, thanks to its own coherent syntax. All the elements contributing to a linguistic query on corpora through the tool are collected (together with their description and an example) in the table 17.31.

Element Tag	Element explanation.	Example
pattern	To start a pattern search.	pattern {N[upos=NOUN]}
without	To start an exclusion pattern search	without {V-[punct]->P; P [lemma="?" "!"];}
{	Where the pattern starts.	pattern {N[upos=NOUN]}
}	Where the pattern ends.	pattern {N[upos=NOUN]}
N	Assigned by the user, to rename the query labels. Can be any letters.	pattern { N [upos=NOUN]}
[	Where the node clause starts.	pattern {N[upos=NOUN]}
]	Where the node clause ends.	pattern {N[upos=NOUN]}
upos	Activation of the selection of the indicated Universal part of speech.	pattern {N[upos =NOUN]}
lemma	Activation of the selection of the indicated lemma.	pattern {N[lemma ="house"]}
form	Activation of the selection of the indicated form.	pattern {N[form ="houses"]}
=	It matches the generic category with the specific query.	pattern {N[form="houses"]}
"	Where the form or lemma starts and ends.	pattern {N[form="houses"]}
	Equal to 'or'.	pattern {X[upos=NOUN ADJ]}
%	To introduce information as notes, metadata, remarks,...	%Query43 pattern {X[upos=NOUN ADJ]}
Person	Changed by any other feature tag. Presence of this feature tag.	pattern {X[upos=VERB, Person]}
Person=2	Changed by any other Person value tag. Presence of this feature-value tag.	pattern {X[upos=VERB, Person=2]}
!	Absence of the feature tag where this element is previously attached.	pattern {X[upos=VERB, !Person]}
<>	Absence of the value tag where this element substitutes "=".	pattern {X[upos=VERB, Person <> 2]}
,	It separates elements inside a node clause.	pattern {X[upos=VERB, Person]}
;	It separates elements between node clauses.	pattern {N[upos=NOUN]; V[upos=VERB]}
<	The node in 1st position immediately precedes the node in 2nd position.	pattern {N[upos=NOUN]; A[upos=ADJ]; N<A}
>	The node in 2nd position immediately precedes the node in 1st position.	pattern {N[upos=NOUN]; A[upos=ADJ]; N>A}
«	The node in 1st position precedes the node in 2nd position.	pattern {N[upos=NOUN]; A[upos=ADJ]; N«A}
»	The node in 2nd position precedes the node in 1st position.	pattern {N[upos=NOUN]; A[upos=ADJ]; N»A}
->	Any dependency between the two implied nodes.	pattern {N[upos=NOUN]; V[upos=VERB]; V->N}
V->	The head of the dependency relation.	pattern {N[upos=NOUN]; V[upos=VERB]; V-> N}
->N	The dependent of the dependency relation.	pattern {N[upos=NOUN]; V[upos=VERB]; V-> N }
V-[subj]->N	A dependency of subject between V as head and N as dependent.	pattern {N[upos=NOUN]; V[upos=VERB]; V-[subj]->N }
V-[~subj]->N	Any dependency except subject between V as head and N as dependent.	pattern {N[upos=NOUN]; V[upos=VERB]; V-[~subj]->N }

Table 17.31: Grew-Match Formalisation Syntax

17.4 Data Management Formalisations

Finally, we must show the different elements used to create the different formalisms in the Data Management section. In this case, we only wanted to represent the operations formally carried out to manage the raw data obtained in UD. The processing of this data is carried out through different formulas. We have tried to follow the classical conventions for the typical and general elements in the syntax of this type of formalization (the use of “ $\wedge$ ” for ‘and’, for example). However, we have decided to create a symbol for other elements to use consistently in the different formalizations. These are explained in the table 17.32.

Symbol	Value
$\bar{x}$	Average number of occurrences of an element in a language
O	Occurrences in a corpus of a language
D_L	Dominant type of a given language L
e	Occurrences of one of the possible type options
E	Set of possible type options
$F(e)$	Frequency of an e in its set E
O_E	Total occurrences of the whole E
P_e	Percentage of occurrences of a type within its element set
T	Fulfillment of the Universal in a language L
W	Not fulfillment of the Universal in a language L
C_U	Percentage of the accomplishment of a Universal
i	imaginary number
n	number of elements
$\sum_{i=1}^n$	Sum of all the n elements starting at 1
$\in$	‘in’ (an element belonging to a set)
$ $	conditional ‘if’ within a set
$\geq$	is greater or equal to
max	the maximum value
$\setminus \{e\}$	excluding the $\{e\}$ element
$\nexists$	no existence of an element
$\Rightarrow$	logical conditional ‘if...then’
$\{$	to include the possible cases
$\wedge$	and
$\vee$	or

Table 17.32: Data Management Formalisations

18

Decisions in Formalisation

Contents

18.1	Basic Word Order	245
18.2	Adpositions	246
18.3	Subjects, Objects, and Verbs	250
18.4	Adjectives	252

It should be remembered that Greenberg’s Universals have been formulated through a natural language, specifically in English. Since we want to conduct a computational check, we must formulate such Universals in a formal language that is feasible for analysis. In addition, the formalization of Greenberg Universals can also help to disambiguate some structures or labels that using a natural language may influence.

The first element that is controversial and transversal in many Universals is what is known as the basic word order.

18.1 Basic Word Order

As is well known, post-Greenberg syntactic Typology bases many of its claims on the linguistic type of languages concerning their subject, verb, and object order. However, as it appears in Mithun (1992), there are different ways of arriving at the basic word order of a language (in order to be then able to ascribe it to an implicature or not), such as purely quantitative, the most neutral or least marked, etc.

Greenberg (1963) defines basic word order as three fundamental aspects of languages that condition their linguistic type and other distributions. For him, these three are:

The first one of these is the existence of prepositions as against postpositions. These will be symbolized as Pr and Po, respectively. The second will be the relative order of subject, verb, and object in declarative sentences with nominal subject and object. Most languages have several variant orders but a single dominant one. Logically, there are six possible orders: SVO, SOV, VSO, VOS, OSV, and OVS. Of these six, however, only three normally occur as dominant orders. The three that do not occur at all, or at least are excessively rare, are VOS, OSV, and OVS. (...) The third basis of classification will be the position of qualifying adjectives.

18.2 Adpositions

Although Greenberg does not mention it in this definition because he takes it for granted, throughout the Universals in which prepositions or postpositions (or adpositions) appear, one can see what he means by preposition and postposition. For Greenberg, a preposition is an adposition placed before a noun with a dependency relation. Accordingly, a postposition is an adposition placed after a noun with a dependency relation.

This implies that when we want to represent Greenberg’s Universals, we should refer only to the relation between adpositions and nouns as a paradigm of prepositions and postpositions without considering other parts of speech.

The first thing we place in the formalization we propose is the title we give to it (in natural language) so that it is more easily identifiable. This goes in the first line of the formalization and is preceded by the symbol “%”. In the case of the preposition, we propose:

$$\boxed{\%Prep} \quad (18.1)$$

Then, on the next line, we start the actual formalization. This must be headed by the word “*pattern*”. The pattern to be searched for always appears inside the symbols “{” and “}”. If we were looking for an element X, we would have the structure:

$$\boxed{\begin{array}{l} \%Prep \\ pattern\{X\} \end{array}} \quad (18.2)$$

Now, we must replace that element X with what is really a preposition. The first of its characteristics is that it is an adposition. The label used in Universal Dependencies for it is “ADP”. This

is its part of speech, whose label in Universal Dependencies is *upos* (Universal part of speech). This information is contained within the symbols “[” and “]”. Before starting with “[”, the contained element must be named. In this case, as it is a simple formalization where we only reflect that the part of speech of the word must be an adposition, we can keep the same label as in Universal Dependencies (ADP). In other cases, for example, if we formalize exclusively feminine adjectives, we could rewrite some tags such as ADJF. However, this outer word is at the discretion of the researcher. In our case:

$$\begin{array}{l} \%Prep \\ pattern\{ADP[upos = ADP]\} \end{array} \quad (18.3)$$

Once we have activated the part of speech corresponding to the preposition, which is the head of the dependency (in the SUD annotation scheme), we detail the dependency we are looking for. First, we use the element “;” to separate the dependency from the preceding. Once it is delimited, we locate the name given to the previous element, which coincides with the part of speech, ADP. Then, we make use of the symbol “–”. Then, we add another box with the symbols “[” and “]”, and inside, we place the type of dependency we want. In the case that we have commented on, following the SUD annotation scheme, it is a dependency of “*comp : obj*” since it is argumental. Once the box with the dependency is closed, we add again the symbol “–” and also the symbol “>”; this is equivalent to the graphical representation of an arrow. At the end of the arrow, we add the label of the dependent, which we have not yet formalized but which, in the next step, we will name “*N*”. There is a separating space between the words and the date behind the “*N*”. In short, we must write:

$$\begin{array}{l} \%Prep \\ pattern\{ADP[upos = ADP]; ADP - [comp : obj] - > N\} \end{array} \quad (18.4)$$

Once we have indicated the dependency we are looking for, we add another “;” to separate the next element, and we start to formalize the characteristics of the dependent. In this case, we only want to restrict the part of speech of the dependent, which must be a noun (without further specifying its characteristics). However, in Universal Dependencies there is a differentiation between NOUN (common nouns) and PROPEN (proper nouns). Since Greenberg did not make this distinction, we assign to our element “*N*” which can be either of these parts of speech. To do this, we start by locating the “*N*” and start a new box with “[” and “]”. In it, we again indicate that we are looking for a “*upos*” equivalent (=) to NOUN or a PROPEN. The formalism for indicating

it is adding “|” between the affected tags (*NOUN|PROPN*). Therefore:

$$\begin{array}{l} \%Prep \\ pattern\{ADP[upos = ADP]; ADP - [comp : obj] - > N; \\ N[upos = NOUN|PROPN]\} \end{array} \quad (18.5)$$

The last aspect, after adding “;” again, that we must characterize the preposition is its order of appearance. This is what will differentiate it from a postposition. This is done by using “<<” between the two elements we have named (*ADP* and *N*). That direction indicates precedence. The opposite direction would indicate the reverse (>>); we would look for postpositions. Using a single “<” symbol instead of two would indicate immediate precedence and not just precedence. In this case, we prefer to mark precedence from a broader point of view, as the opposite would not allow us to capture occurrences where there is an article between the preposition and the noun, for example. That is, the complete formalization of a preposition would be (18.6):

$$\begin{array}{l} \%Prep \\ pattern\{ADP[upos = ADP]; ADP - [comp : obj] - > N; \\ N[upos = NOUN|PROPN]; ADP << N\} \end{array} \quad (18.6)$$

And a postposition would be (18.7):

$$\begin{array}{l} \%Post \\ pattern\{ADP[upos = ADP]; ADP - [comp : obj] - > N; \\ N[upos = NOUN|PROPN]; ADP >> N\} \end{array} \quad (18.7)$$

However, we believe that, perhaps, if we slightly modify this definition of adposition, we may be able to capture new phenomena, which could contribute to the nuance of the Universal (to extend it, to restrict it, to confirm it, to disprove it, etc.). Thus, once the Universal *à la Greenberg* has been formalized and the data have been extracted in the most representative way, we will also offer alternative formalizations (and analyses) that can be interesting. In the case of adpositions, for example, we can see how this part of speech behaves when its *comp : obj* dependent is not a noun but a pronoun or a verb.

Moreover, we can present a more general and loose definition of *ADP*, which admits any part of speech as dependent on this argument relation (*comp : obj*) of an *ADP*. This would make us also consider less prototypical and, therefore, less frequent and more borderline relations, but no

less valid from the point of view of grammar. For example, to take into account any part of speech in prepositions, we propose the pattern (18.8):

$$\begin{aligned}
 &\%Prep - ANY \\
 &pattern\{ADP[upos = ADP]; ADP - [comp : obj] - > ANY; \\
 &ANY[upos = ADJ|ADV|INTJ|NOUN|PROPN|VERB| \\
 &ADP|AUX|CCONJ|DET|NUM|PART|PRON| \\
 &SCONJ|SYM]; ADP >> ANY\}
 \end{aligned}
 \tag{18.8}$$

It is well known that Typology has traditionally based its data on grammars, where the main and most representative cases of a language appear, for example, the position of adpositions concerning nouns. We do not know if we can extend this same order to adpositions with adjectives, numerals, determiners, etc. If the hypothesis is confirmed, we can extend the results to any part of speech, which is not possible with grammars. If the hypothesis is not confirmed, we could detect and show some anomalous behavior of the part of speech in specific languages.

We are fully aware that perhaps there will be some parts of speech that do not present any kind of dependency with *ADP*; however, we wanted to present all of them for two reasons.

On the one hand, we do not know if, in any of the thousands of languages in the world, there is any particular construction between an *ADP* and a part of speech that in a Western European language would be unthinkable. Although we may suspect that, for example, we will not find dependencies between *ADP* and *PART*, it may be that in some languages, it is possible and frequent. Alternatively, we can think of a relation of *ADP* and *SYM* in examples like “in \$”.

On the other hand, if we want to make a theoretical formalization about a possible definition of adposition, we must represent any possible part of speech representing the phenomenon. Therefore, we have included all POS except “*X*” and “*PUNCT*”. The first one is because it is a wildcard tag used to tag unclassifiable problems, such as the presence of “xfgh” (example in English). The second one is because the punctuation always receives a dependency, also called *punct*, from the head of the whole sentence, so it will never appear.

To sum up, we want to not only represent and offer quantitative data on real utterances representing Greenberg Universals but also to represent further more specific, complex, and fine-grained results regarding less studied (but equally important) situations.

18.3 Subjects, Objects, and Verbs

Referring to these three essential elements of sentence syntax, we must strongly emphasize a fundamental aspect of Greenberg’s definition: “the relative order of subject, verb, and object in declarative sentences with nominal subject and object.”

1. We must consider the presence of the three elements simultaneously in the analyzed sentences. We must analyze all the sentences containing these three elements to do this.
2. We must only take declarative sentences, so we must filter sentences with interrogations and exclamations, for example, which can alter the order of these elements.
3. We must only consider direct objects and subjects that are nouns, that is, that their part of speech is *NOUN* (common noun) or *PROPN* (proper noun).

This third element is the most controversial point when approaching Greenberg’s proposal. Other researchers have already tried to analyze the occurrences of Word Order in the world’s languages as proposed by Greenberg (Choi, Guillaume, & Fort, 2021; Choi, Guillaume, Fort, & Perrier, 2021; Yan & Liu, 2023). Their proposal includes three elements: *NOUN*, *PROPN*, and *PRON*. We wonder why they added *PRON* to the *à la Greenberg* analysis (and why not add others then). We believe that it is due to the confusion of believing that the terminology used by Greenberg and the one currently used are equivalent when this is not the case since we are in different historical moments and frameworks.

Choi, Guillaume, and Fort (2021), Choi, Guillaume, Fort, and Perrier (2021) and Yan and Liu (2023) have simply applied the distinction that exists in the Universal Dependencies annotation guidelines, which oppose *nsubj* to *csbj*, in other words, nominal subjects to clausal subjects. The nominal subjects correspond to the three categories they analyze in their formalization, as exemplified in the UD annotation guide. However, we understand that Greenberg is not making such an opposition since the most frequent and typical distinction at the time (and always in general) was between nominal subjects and pronominal subjects, for example.

To ensure that our reading is more accurate with what Greenberg intended, we have considered a confirmation criterion.

We have reviewed what happens explicitly in the Universals Greenberg proposed and the examples he offers in them. We do not find any mention of pronouns in the examples of Universals

that he explicitly cites word order. Furthermore, it is much more evident if we go directly to the formulation of some Universal, such as Universal 25, where he explicitly uses the opposition between “nominal object” and “pronominal object”. Therefore, we understand and confirm that the fact that he explicitly mentions nominal objects and subjects means that pronouns are discarded to determine the basic word order of a language. When Greenberg mentions nominal subjects and objects, we understand that clausal subjects or objects are not considered. We believe this methodology we are formalizing is the most correct to represent how Greenberg determined the basic word order of languages by this definition of basic word order. However, as we have already commented with the adpositions, this does not mean we renounce proposing other data obtained through other definitions of basic word order (Mithun, 1992). Once we have formalized the different basic word orders as we believe Greenberg would like (and we have obtained the data in the languages we can), we also propose other, more open formalizations that allow us to obtain other metrics that will be interesting to compare. We will do this, especially in Universal 1, to be able to represent what would be the quantitative differences that would exist between the different definitions of word order. Thus, we could assess which approaches are most desirable. How it is analyzed can lead to significant changes in the categorization of languages, as can happen with Afrikaans, which is a complicated language to simplify due to its peculiar syntactic structure. Depending on the theoretical basis, this may have different implications. Therefore, we will look at the different cases and decide which is the best for dealing, in general, with languages in all Universals. The section corresponding to Universal 1 will offer more details about the methodology used.

In short, the proposal we make to characterize as Greenberg states the OSV order, for example, would be (18.9):

$$\begin{array}{l}
 \%OSV \\
 pattern\{V[upos = VERB]; V - [subj]- > S; S[upos = \\
 PROPEN|NOUN]; V - [comp : obj]- > O; O[upos = \\
 PROPEN|NOUN]; V >> S; S >> O; V >> O\} \\
 without\{V - [punct]- > P; P[lemma = "?"|"!"]\}
 \end{array} \quad (18.9)$$

In this formalization we find, in the first line, the information of what order we are setting. In the next one, the created pattern begins, where the words with part of speech of *VERB* are called *V*. Furthermore, these verbs have a subject dependency relationship with the *S* elements, which we have restricted to any word that is *NOUN* (common noun) or *PROPEN* (proper noun). This same *VERB* is also the head of another dependency, in this case, *comp : obj*, with an *O* object as

it's dependent (which again is any *NOUN* or *PROPN* word). Furthermore, we require that the verb appear after the subject (“ $V \gg S$ ”) and after the object (“ $V \gg O$ ”). We also ask that the object appear before the subject (“ $S \gg O$ ”). In the last line, we add the filter that we do not want to consider any sentence with interrogations or exclamations.

18.4 Adjectives

When Greenberg mentions the order of the adjective as a modifier, it explicitly does it regarding the order between adjectives and nouns. Again, there can be another part of speech modified by an adjective. However, Greenberg focuses on Typology with the most frequent and common orders (the prototypical ones and, thus, the ones cited in grammars). In short, again, we are going to offer a double approach.

On the one hand, we are going to formalize all the Universals related to adjectives by considering them as modifiers of nouns (common and proper) and indicating the restriction of appearing as $A - N$ (adjective before noun) or $N - A$ (noun before adjective), as we show in (18.10):

$$\begin{array}{l} \%ADJ - N \\ pattern\{A[upos = ADJ]; N - [mod]- > A; N[upos = \\ PROPEN|NOUN]; A << N\} \end{array} \quad (18.10)$$

In this case, we name all the adjectives with the label A . And the common and proper nouns with the label N . Then, we set all relations with an N as head, a modification dependency (*mod*), and an A as dependent. In addition, we restrict the order where A appears before N ($A << N$), for example.

On the other hand, we will present other possibilities: what happens when adjectives modify pronouns, determiners, numerals, or any part of speech in general, for example. The main objective of this change is to improve the knowledge we currently have, again. We can either extend the knowledge of adjective-noun correlations to others or, conversely, we can point out that this relationship is unique to the adjective-noun and does not apply to other parts of speech. Either way, we will offer more precision applied to non-prototypical but essential language relations from a usage-based approach.

19

Data Management

Contents

19.1	Sample	253
19.2	Data and Limitations	255
19.3	Script	258
19.4	Data Cleaning	258
19.5	Calculating Universal 1	260
19.6	Implicational Formulae	265

19.1 Sample

The languages used for the claims we offer in this study are all the available languages in UD (version 2.11). That is a convenience sample. The reasons for not working with a balanced representative sample in terms of probability or variety are several.

First, it is impossible to provide a good sample of such characteristics with the current data available in Universal Dependencies. More languages (especially from macro-areas with traditionally lower representation in linguistics) are needed for adequate sampling in terms of probability or variety. From a typological perspective, the balances sought to be respected are related to linguistic, genetic, and areal factors. When following a convenience sample like the one we are using, we encounter a prevalence of specific characteristics based on the data available in Universal Dependencies (UD). We provide a detailed list in the appendices with all the relevant characteristics of the languages available in UD. You can individually consult the data for each language.

However, in general, it should be noted that there is a clear tendency to work with SVO languages, Indo-European languages, and Eurasian languages.

Regarding the linguistic type, you can refer to table 19.1, where we also provide the availability of data from WALS, representing a more appropriate detail regarding these balances (thanks to the 2,662 languages with at least minimal entries). In this case, we see that in UD, 55% of the languages are SVO, while in WALS, they represent 41%. This higher proportion should likely be balanced out in the SOV languages, which account for 28% in UD and 35% in WALS. As for the other word orders (VSO, VOS, OVS, OSV, NDO), there are some differences, but they are minor.

Order	% UD	% WALS
SVO	55	41
SOV	28	35
VSO	5	7
VOS	0	2
OSV	1	0
OVS	1	1
NDO	11	14

Table 19.1: Percentage of Languages Order in the Sample

In terms of language families, Indo-European is the most represented family in the UD set, accounting for 46% of the languages. In WALS, this family represents only 7% of the languages. Although this data is quite striking and indicates the bias that may be present in its composition, it is essential to note that the data in WALS is not as varied as it may seem. While there may be a relatively low percentage of Indo-European languages in WALS (176 out of 2,662 languages), this number can be misleading. Many of these languages in WALS only have a few described linguistic features, or even just one. This means that in the case of the example we took, Universal 1, it may not be explicitly described in those 2,662 languages but in half of them. The languages that are not described, for the most part, are non-Indo-European languages, as the lack of data and descriptions applies to the same languages as in UD.

Regarding macroareas, we should mention both a positive and a negative aspect. The positive aspect is that all macroareas are represented (Eurasia, Africa, Australia, Papunesia, North America, and South America). However, the majority of languages belong to Eurasia (76%). We believe that the representation of South America (9%) and Africa (7%) may be sufficient for reliable results. However, we believe that in the future, it is necessary to have more languages, families, and genera from Australia (1%), Papunesia (4%), and North America (1%). Therefore, we think it would be beneficial for the tool to focus its efforts on working with languages from these three macroareas.

Secondly, we believe that this study is relevant to the data offered in terms of Universality and to the individual typological representation of the languages. In our case, we are offering the typological representation quantitatively of some languages for which we do not have this data. Therefore, we believe it is crucial to have such information in as many languages as possible as an initiatory work, without vetoing from the analysis any language for which we can offer new data or data that can be contrasted with data previously presented by another author.

Thirdly, in their study, Guzmán Naranjo and Becker (2022) mathematically show how it is not necessary to calculate any special sampling (of variety or probability) if the sample of languages is sufficiently large. Although the current number of languages analyzed could be much larger, we believe it is worthwhile to lay the groundwork for such a study, which, if possible, can provide much more reliable data by adding new treebanks without affecting the calculation already done.

Fourthly, the number and variety of languages analyzed allow us an adequate (and coherent) representation of the theoretical proposal or innovation we present in this doctoral thesis. The data we handle show a reliable trend in general terms, which could be slightly modified in exact quantitative terms. However, we are not concerned with quantitative detail (as we have already argued) because of the possible bias to which any available data is subject: we believe that what is relevant is to know the general location or representation of the phenomenon.

19.2 Data and Limitations

The treebanks available in UD 2.11 are 241. As mentioned above, all of them will be taken into account. This corresponds to the analysis of 143 different languages. In WEIRD languages, as in the grammar approach, we usually have multiple corpora of different typologies, which contribute to obtaining a much broader and more representative picture. In the case of minority and traditionally marginalized languages, we only have one corpus if we are lucky. Moreover, in these cases, the corpus is usually conditioned by a specific theme, such as religious or literary texts. However, we can do nothing to mitigate this, as these are the only (or some of the few) data available for these languages in other linguistic approaches. For example, the Old Slavic sample has very clear religious overtones. However, this is the only information we have for such a language in general, as these texts have survived. We must remember that we will work with the SUD annotation scheme and not UD, as we have already developed previously.

This undesirable situation forces the researcher either to give up conducting such a study or to warn of the problem and carry it out anyway. In this case, we have opted for the second option.

Regarding the amount of data available, we must also mention quantitative limitations. Some corpora have millions of sentences or words (and others with only a few thousand). Given the multiple formalizations we propose, the large number of operations required makes it impossible from a technical point of view to study all possible sentences in all corpora. With the hardware available to the researcher, the calculation would take days (if it does not collapse first). Therefore, given the novelty of the proposed study, we consider a valuable and representative approach to the first 1000 sentences of each of the available corpora. This implies the analysis of a number close to 241,000 complete sentences (with all the morphosyntactic tokens of each word).

Multiplying the number of sentences analyzed by the number of patterns we have created to check different aspects, we obtain a total number slightly lower than 243 million pattern/sentence operations.

Moreover, this reduced selection of sentences allows us to homogenize the results between corpora much more and to obtain more easily comparable and fair data.

Some researchers may criticize the fact of selecting the first 1000 sentences and not 1000 sentences randomly from the different corpora. The criticism usually occurs because there may be some bias taking such sentences. Depending on the type of corpus, these sentences may be special. For example, in the case of literature, it may be that the introductions or first chapters have characteristics that do not appear equally consistent in the rest of the work. Furthermore, we also understand that the bias occurs not because they are the first sentences but because of the type of text that is itself biased (e.g., a novel). However, in the corpus Typology we work with, in most cases, it does not matter whether the sentences are first sentences or not, as they are collections of news items, wiki comments, etc. For a complete list of corpus typologies, the corpus name, and the language analyzed, please refer to the annex documents. In addition, more details on the different corpora can be accessed directly through Nivre et al. (2023).

However, we believe this risk is quite marginal and does not affect the results much (especially if we do not want to consider exact quantitative data but a classification by means of the Fuzzy Logic model). The reason for taking these first 1000 sentences and not randomly is falsifiability. We believe that in works of this type, it is essential to guarantee the provenance of the data and to facilitate their traceability. Any researcher who wants to reproduce or review the data knows perfectly well on which sentences these claims are based. In the case of providing a random

selection, even if the analyzed sentences could be offered, the process would be equally deficient. On the one hand, reviewing from a computational point of view would be more complicated since the corpus is always handled in its entirety. On the other hand, it would not be possible to justify the truly random nature of the data since there can always be a suspicion that the sample has been adulterated at the researcher’s convenience. In summary, we believe this choice is the most interesting and ethically responsible from a scientific and objective point of view.

The maximum number of 1000 occurrences (and no other) has been chosen because it is considered to be the number that allows the greatest number of occurrences and variation without being too high to give uneven results between very small corpora and massive corpora. It is an excellent unifying limit.

Although we believe the results between a sample of 1000 sentences and a sample of millions of sentences do not vary much (especially in the Fuzzy Logic terms used in this thesis), we wanted to check this behavior. For example, we can select any aspect, such as the distribution of pronominal direct complements, to compare the behavior of such a phenomenon with the limit set and without limit, as shown in table 19.2. In order to select the example languages shown, we have listed some of the languages and their results based on alphabetical order.

Language	Without Limit	1000 Limit	% Without Limit	% 1000 Limit
Danish	High / VO	High / VO	77	77
Dutch	Low / OV	Low / OV	27	29
English	High / VO	High / VO	86	84
Erzya	Medium /NDO	Medium / NDO	34	35
Estonian	Low / OV	Low / OV	32	28
Faroese	High / VO	High /VO	86	85
Finnish	Medium / NDO	Medium / NDO	46	45
French	Low / OV	Low / OV	14	12
Frisian Dutch	Medium / NDO	Medium / NDO	45	45
Galician	Low / OV	Low / OV	20	24

Table 19.2: Comparison between 1000 Sentences Limit or without Limit

The results are generally the same. There is no label that varies. Nor are there any major changes in the exact percentages; they tend to coincide or vary by 1 or 2% in most cases. Therefore, we can be more confident that our proposal can be extrapolated to the totality of the sentences in

the different corpora.

19.3 Script

In order to obtain the results of the treebanks indicated (all those available) about the patterns we have been creating to capture the different linguistic characteristics, we must create a script in an orderly manner that allows easy reading and management of the data.

To do this, on the vertical axis, in each line, we have placed each available treebanks (241) in column B, preceded by the number of the Universal being processed in column A. Once finished with the first Universal, this list is repeated with the number of the following Universal. In line 2, we will see in column A “1” and column B “sud_abaza-atb”, which corresponds to the different data related to Universal 1 appearing in the Abaza language’s treebank. Once the 241 treebanks are completed, line 243 will show in column A “2” and column B, again, “sud_abaza-atb”, corresponding to the verification for the same language of different aspects of Universal 2.

On the horizontal axis, in the columns, after the information discussed above in columns A and B, we have the text “Pattern 1” in column C, “Pattern 2” in column D, and so on up to column BW, which corresponds to “Pattern 73”, which is the maximum number of patterns we have created to work on a single Universal.

Each pattern number is assigned to each of the formalizations we are looking for, which we have in the search list in the created plot. In column C (pattern 1) of line 2 (Abaza), we will get the result of occurrences of the basic order SVO according to Greenberg in such a language. In column D (pattern 2) of line 2 (abaza), we will get the number of SOV occurrences. On the other hand, in columns C (pattern 1) and D (pattern 2) of line 243, we will obtain the number of occurrences in Abaza of adposition + noun and noun + adposition, respectively.

19.4 Data Cleaning

Once we have run the script created with the over thousand patterns we want to review in the 241 treebanks available and obtain the raw results, we must organize our data.

First, we convert the script into small independent files, each corresponding to a Universal. In this way, the data handling is better and more organized. In our case, we have converted the

results to a .xls file. We have as many documents as Universals we want to check (and we assign the name of the Universal to each file).

The next step is to delete the Universal number column (used to create these independent files), which now has no use. This column can also be used to not alter the order of the components or to add other information, such as the original size of the corpus (although the results are limited to 1000 sentences at most).

Subsequently, we replace line 1 with the line of the patterns we have searched for. That is, we change, for example, “Pattern 1” by the pattern that applies in the case of “Pattern 1” of Universal 1. Once this change is made, the formalization of each pattern is eliminated. We are left with only the general information or the title of each pattern (as “%1-SVO”, for example), and we eliminate the formal mark “%” and the number of the Universal. In other words, the final result will be that in the “Pattern 1” cell, we will now have “SVO”, and in the “Pattern 2” cell, we will have “SOV”.

Later, we merge the results from different treebanks corresponding to the same language. For example, in the case of Spanish, we add up the SVO occurrences from the “sud_Spanish-ancora”, “sud_Spanish-gsd”, and “sud_Spanish-pud” corpora, and then divide the result by 3. We perform this action for all treebanks that share the same language (e.g., in this case, all of them starting with “Spanish”). We do not apply this step to Chinese corpora. In this case, even though they share the same label, the linguistic varieties are different enough for such merging to be artificial and misleading. There can be important differences between Chinese corpora, as can be observed when comparing the GSDsimp (Simplified Mandarin), HK (Hong Kong variety), and Kyoto (Classical Chinese) corpora. Cantonese is already labeled as an independent language in UD. This calculation will respect the proportions and provide the result, considering all the treebanks in a unified manner.

To merge and homogenize the results, we propose the formalization we have created in equation (19.1). We have changed the name of the languages affected by the merge from lower to upper case to make the phenomenon easier to read.

$$\bar{x} = \frac{\sum_{i=1}^n O_i}{n} \quad (19.1)$$

Where:

- $\bar{x}$: the average number of occurrences of an element in a language.

- O_i : stands for the occurrences in the “ i ” corpus of a language.
- n : stands for the number of corpora in a language.

That is to say, in other words, if we had three different corpora, the first two of which had five occurrences each and the last one 11 occurrences, the final result would be:

$$\bar{x} = \frac{5 + 5 + 11}{3} = \frac{21}{3} = 7 \quad (19.2)$$

19.5 Calculating Universal 1

Once we have sorted and cleaned the documents, we need to know how to extract the information we are interested in from the occurrences obtained from the created patterns. This process is similar for all Universals, although there may be slight differences depending on whether or not many operations need to be performed, depending on the complexity of the Universal originally formulated by Greenberg. We will limit ourselves to explaining step by step how the Excel formulas were applied and how the raw data were managed and converted into more representative data in the case of Universal 1. We believe that by understanding this example, you can grasp the functioning of the rest of the formulas for the other documents, which are attached as appendices to this thesis (the formulae and the results).

In the patterns about basic word order, we find the results for the six possible word orders for a sentence with the three elements proposed by Greenberg. However, in the following patterns, we find the same results applied only to pronouns, followed by the sum of common nouns, proper nouns, and pronouns. Then, without the parts of speech restriction. Subsequently, we obtain the data for the opposition between SV/VS and OV/VO separately, in nominal terms, followed by pronominal terms, then with the sum of common nouns, proper nouns, and pronouns, and finally, without the parts of speech restriction.

For a better visualization of the labels in the .xls file, we propose adding a separate column between the different classifications. This column will contain the distinguishing information, such as “Pronoun Word Order”, while the individual columns will only contain linguistic type information. In other words, we convert pattern 7, which was “SVOpn”, into “SVO” because we have added the information “Pronoun Word Order” in the previous column. This applies to all classifications.

We have also added the freezing of row 1 and the language column, facilitating data browsing and the .xls management. We have also included the ability to filter in row 1, allowing easy filtering and data analysis.

Later, we added surveys to gather information about the dominance of each of the aspects we want to review, such as Basic Word Order Dominance, Pronoun Word Order Dominance, Pronoun + Noun Word Order Dominance, Any PoS Word Order Dominance, Basic Word Order Dominance SV, Basic Word Order Dominance VO, Word Order Dominance SV pron, Word Order Dominance VO pron, Word Order Dominance SV pron + noun, Word Order Dominance VO pron + noun, Word Order Dominance SV Any PoS y Word Order Dominance VO Any PoS.

Subsequently, we must calculate the dominant order of each language, as Greenberg indicated. We will follow the methodology of Gries (2013) to do it: a dominant element is the first ranked element that exhibits a value equal to or greater than twice the second maximum value of a set. Otherwise, there is no dominant order, just one more frequent than the other. In order to perform such a check, we must carry out several steps. We have decided to create a formalization that captures this rule in a mathematical way (19.3):

$$D_L = \begin{cases} e \in E \mid F(e) \geq 2 \cdot \max(F(E \setminus \{e\})) \\ e \nrightarrow \Rightarrow \text{NDO} \end{cases} \quad (19.3)$$

Where:

- D_L : represents the dominant type of a given language L .
- e : represents the occurrences of one of the possible type options.
- E : stands for the whole set of possible type options.
- $F(e)$: stands for the function of the frequency that the element e has in the set.
- NDO: stands for “nondominant order”.

That is, we must read the formalization we have created as:

The dominance of a type in a given language L will be equal to the type of an element e within a set of elements E that exhibits a frequency of occurrences equal to or greater than twice the maximum value of the set of elements E excluding the analyzed element e itself. If such a value

exists, the result will be that value. If no such value exists, the type dominance result will be NDO. That is, when we say that Spanish is a language with a dominant word order SVO, we have observed that the occurrences of sentences with SVO structure (with the limits marked by Greenberg) appear twice as many times or more than the second most observed basic order.

We need to apply this formalization to each of the proposals we have discussed earlier (with different combinations of parts of speech restrictions). All of this will give us a clear overview of the behavior of each language (in the available corpora) regarding the investigated linguistic type.

After carrying out such a check, we must calculate the total universe of the different occurrences of the different elements corresponding to a category to determine whether the sample is sufficiently representative to make the claim we are making. For this purpose, we have decided to summarize this will in the formalization (19.4):

$$O_E = \sum_i e_i \quad (19.4)$$

In this case, this formula is read as:

O_E (the occurrences of an element) is equal to the sum ($\sum_i$) of the occurrences of the different possible types of manifesting that element e_i .

Once we have this data, we can carry out the checks that show the significance of the sample and will help us to determine what threshold we should set. We should mention that if the sample is not considered representative, the result will be considered not applicable. This is done because we do not consider these small results as negative; we simply refuse to value them and take them as occurrences (in a positive or negative sense) for our study. In other words, with this consideration, we make the data for this language look as if they do not exist. We are waiting for an improvement in the representation of the linguistic data in order to be able to offer more accurate claims.

The significance check has been carried out automatically thanks to Python, as shown in the example in figure 19.1



Figure 19.1: Example of the Checking of Significance of our Data

We have carried out multiple experimental tests, checking different weights, which have allowed us to obtain thresholds that allow us to eliminate the cases with very low occurrences and that can clearly show a biased representation of the phenomenon but at the same time be as flexible as possible to be able to characterize with relative confidence the largest possible number of languages in the world. In other words, thanks to these tests, we have been able to determine which are the numbers that allow sufficient reliability, but at the same time being as low as possible to be able to encompass the maximum possible number of languages and thus make the tested set as large as possible. Considering the dominance criteria we use, which require a high proportion to assign the corresponding label, we have found that we can obtain valid results with a universe of more than ten occurrences.

This minimum threshold for not ignoring the results obtained from a language is, we believe, a more realistic option than setting a minimum corpus size to be taken into account. We believe that the minimum number of tokens a corpus must have to yield reliable results may vary depending on the phenomenon being analyzed. If the Universal proposed by Greenberg deals with aspects such as the position of the adjective modifying a noun or the position of the adposition about a noun, we believe that corpora with a low number of tokens can be equally helpful since these are phenomena that can appear multiple times in the same sentence. On the other hand, in these same corpora, Universals referring to much more specific aspects and with a lower frequency of use will not yield representative results. Therefore, to not constantly modify this minimum threshold of representativeness, we have opted for a count of occurrences of the phenomenon.

Later, we can also perform quantitative calculations (or calculations without labels). Although these may not be directly useful for this particular Universal, they can be carried out for two

reasons. Firstly, it may be interesting data for scientists who do not work with categories (because they do not want to establish any comparison with Greenberg) and simply want to have quantitative data. This occurs within the Typometrics approach. Secondly, although it may not be the case for this specific Universal, there are instances where Greenberg mentions the existence or emergence of alternative structures as a legitimizing or prohibiting element of a Universal. In order to determine that a language has alternative structures (i.e., non-dominant but existing), we need to take a quantitative approach that sheds light on such data.

For the quantitative calculation, first, we need to obtain the total universe of the object of study. We already have this information from the formalization (19.4).

Then, we must determine the proportion of occurrences of the different orders within that universe. To do this, we need to calculate the percentage of each possible type of the studied element, which can be summarized in the formalization (19.5):

$$P_e = \frac{e}{O_E} \cdot 100 \quad (19.5)$$

In this case, to obtain the result of the percentage of each type (P_e), we take advantage of previous elements. Therefore, this formalization can be read as the percentage of a type (P_e) is equal to the occurrences of the type being analyzed (e) divided by the total occurrences of the element encompassing all types (O_E) multiplied by 100.

Next, we can determine whether the Universals proposed by Greenberg hold in the different languages of our set. First, we have to discard or filter out all the cases in which the result is not applicable for different reasons (the number of occurrences does not exceed the significance cut-off, the corpus of the language presents some error in some of the elements analyzed, etc.).

Then, we have to check which languages this Universal holds for. The values that make a Universal true in a language change in each Universal. In the case of Universal 1 that we exemplify, we understand what we show in formalizations (19.6) and (19.7):

$$D_L = \text{SVO} \vee \text{SOV} \vee \text{VSO} \Rightarrow T \quad (19.6)$$

$$D_L \neq \text{SVO} \vee \text{SOV} \vee \text{VSO} \Rightarrow W \quad (19.7)$$

The type of formalization we have presented solves in a rather strict (and clear) way the problem related to languages with nondominant order. In the formulated Universal, Greenberg simply highlights common languages that exhibit a dominant order in which the nominal subject precedes the nominal object when they co-occur (SVO, SOV, VSO). Therefore, to be strict, we understand that a language whose order is other than the one formulated above (nominal objects preceding the subject or not having a dominant order) does not fulfill the Universal.

Once we have classified the languages as T (true) or W (wrong), depending on whether they fulfill the Universal or not, we must add up all the cases in which we find T and divide by the total universe of languages to which we apply the Universal (positively or negatively). This is summarised in the formalization (19.8):

$$C_{U_i} = \frac{\sum_{i=1}^n T_i}{\sum_{i=1}^n T_i + \sum_{i=1}^n W_i} \cdot 100 \quad (19.8)$$

Where:

- C_{U_i} : stands for the weight of the realisation of a Universal i .
- $\sum_{i=1}^n$: stands for the sum of all elements that exhibit “T” or “W”.
- T_i : stands for each of the T .
- W_i : stands for each of the W .

That is, this formula should be read as:

The compliance weight of a given Universal (W_{U_i}) is equal to the sum of all T -valued languages divided by the sum of all T -valued languages plus the sum of all W -valued languages multiplied by 100.

19.6 Implicational Formulae

Next, we show what other extra basic operations have to be carried out when we are in front of an implicational Universal and want to check the correlation between those variables since it was not necessary to do this in Universal 1.

In cases like Universal 3, where we need to determine if a language from a specific type is predominantly prepositional or postpositional, we simply need to use a formulation similar to the one presented in (19.3).

Later, since we have an implication, additional steps are involved in reviewing the accomplishment of the Universal in different languages. This is summarized in the formulations (19.9) and (19.10).

$$D_L = VSO \wedge \text{prep} \Rightarrow T \quad (19.9)$$

$$D_L = VSO \wedge (\text{post} \vee \text{NDO}) \Rightarrow W \quad (19.10)$$

Where:

- VSO: stands for the basic dominant word order of Verb + nominal subject + nominal object.
- prep: stands for the basic dominant word order of adpositions before the dependent noun.
- post: stands for the basic dominant word order of adpositions after the dependent noun.

Therefore, these formalizations are interpreted as:

If the dominant type of language L is VSO and prepositions, then the result will be satisfactory (T). If the dominant type of a language L is VSO and (postpositions or does not exhibit a dominant order), the result will be unsatisfactory (W).

It should be noted that any logical possibilities not contemplated in this formalization have been previously discarded (as “not applicable”), and the cases discarded for other causes, such as errors in the corpus annotation or insufficient occurrences to trace a claim in that language. In other words, all languages that do not exhibit a dominant VSO order have been discarded in this particular check because it is a language type not covered by the Universal implication.

Obviously, with each subsequent check in which we want to assess the behavior of other similar structures (e.g., adpositions without the part-of-speech constraint of the dependent) or in the linguistic types of other Universals, the labels referred to will have to be changed. Depending on each case, more elements should also be added (by means of $\wedge$ or $\vee$) in the Universals that involve a more complex structure, including more types simultaneously.

In the following results section, we want to offer the data in the most straightforward and cleanest possible way to facilitate the reading. The intermediate data obtained are retrievable and verifiable in the file corresponding to each Universal in the appendix.

UNIVERSITAT ROVIRA I VIRGILI

GLU - A GRADIENT MODEL FOR LANGUAGE UNIVERSALS. A COMPUTATIONAL REVISITATION OF GREENBERG UNIVERSALS

Antoni Brosa Rodríguez

20

Universals Classification According to the Tool

Contents

20.1	Universals through Typometrics	271
20.2	Universals through Grew-Match	272
20.3	Universals through UD Treebank Hub Page	273
20.4	Universals through Combination of Different Methods	274
20.5	Universals with Current Limitations	274

Different ways of processing or obtaining UD information exist to better understand Greenberg Universals. The following pages provide a classification of the Greenberg Universals regarding the methodology that can be used and that we will use in the next chapter.

If the information is available in UD, all universals can be checked through Grew-Match. However, when available, we decide to use different methodologies for two main reasons. First, we want to obtain results in a different output or easier way for a linguist (for example, more visual thanks to Typometrics). Second, one of the secondary goals of this study is to show how a typologist can benefit from resources in Computational Linguistics.

Then, we want to show all the possible paths to reach information, explain them, and show the easiest way for a linguist to get the data. When a technique different than Grew-Match is used, we also offer the Grew-Match option, for comparability and scientific purposes. A few Universals can be checked through Typometrics (the most basic and simple ones). Also, some of them can be extracted directly from the Treebank Hub Page, that is, from the website that each corpus has

in the Universal Dependencies website as a presentation. Some of them need a combination of different methods, even using other tools, such as PanLex, as explained in the next sections. To our knowledge, this kind of classification we propose is original.

Universal	Checkable in UD	High-Universal	Possible modification
1	YES	YES	Checking other PoS (not fruitful)
2	YES	NO	Postpositions
3	YES	YES	Adposition with any PoS
4	YES	YES	SVO & Prep
5	YES (only partially)	YES	Opposite direction of genitive in SOV
6	YES (but scarce data)	YES	Without nominal restriction
7	YES	YES	Other SOV & SVO combinations
8	NO		
9	NO		
10	NO		
11	YES (only partially)	YES	Question word initial
12	NO		
13	YES	NO	VO & depending verb after
14	YES	NO	Typometrics, SVO + adv clause, SOV + adv clause
15	NO		
16	YES	YES	SVO, SOV, SOV without inflection restriction
17	YES	YES	No nominal restriction, SOV, SVO & adjectives after
18	YES	YES	Demonstratives, numerals, definites, indefinites
19	YES	NO	
20	YES (only partially)	YES	Separated 3 elements
21	YES	NO	Adverbs after & VO
22	NO		
23	YES	NO	Adposition & proper noun
24	YES	YES	Relative precedes noun & adjective precedes noun
25	YES	YES	Verbal regimes
26	NO		
27	NO		
28	NO		
29	NO		
30	YES	NO	Only mood
31	YES	YES	
32	YES	YES	
33	NO		
34	YES	YES	Other numbers
35	NO		
36	YES	YES	Other categories
37	YES	YES	
38	NO		
39	NO		
40	YES	NO	Inflectional categories one by one
41	YES	YES	
42	YES	YES	Person & number
43	YES	YES	Other PoS
44	YES	YES	Persons one by one
45	YES	YES	Other PoS

Table 20.1: Summary of Tested Universals, their Value and Modifications

In table 20.1, we have offered a summary in which we show the Universals that are testable through UD and those that are not, whether we have been able to validate them as High-Universal or not and the possible modifications or extensions that we will make to the Universal.

20.1 Universals through Typometrics

In this group, we generally find Universals that contain only a single dependency in each part of the Universal. These Universals cannot contain any mention of any morphosyntactic feature, which cannot be mentioned in this mode. They can be checked through Typometrics (and Grew-Match):

- **Universal 1:** In declarative sentences with nominal subject and object, the dominant order is almost always in which the subject precedes the object.
- **Universal 3:** Languages with dominant VSO order are always prepositional.
- **Universal 4:** With overwhelmingly greater than chance frequency, languages with normal SOV order are postpositional.
- **Universal 13:** If the nominal object always precedes the verb, then verb forms subordinate to the main verb also precede it.
- **Universal 16:** In languages with dominant order VSO, an inflected auxiliary always precedes the main verb. In languages with dominant order SOV, an inflected auxiliary always follows the main verb.
- **Universal 17:** With overwhelmingly more than chance frequency, languages with dominant order VSO have the adjective after the noun.
- **Universal 19:** When the general rule is that the descriptive adjective follows, there may be a minority of adjectives that usually precede, but when the general rule is that descriptive adjectives precede, there are no exceptions.
- **Universal 21:** If some or all adverbs follow the adjective they modify, then the language is one in which the qualifying adjective follows the noun and verb precedes its nominal object as the dominant order.
- **Universal 24:** If the relative expression precedes the noun either as the only construction or as an alternate construction, either the language is postpositional, or the adjective precedes the noun or both.

- **Universal 25:** If the pronominal object follows the verb, so does the nominal object.

20.2 Universals through Grew-Match

Grew-Match allows the formalization of any information that can be obtained in UD. All Universals that can be analyzed with other tools, which we have prioritized as more accessible or simpler for a linguist, can also be analyzed. The only Universals that cannot be analyzed with Grew-Match are those for which UD does not provide annotation, i.e., about intonational patterns or the morphological structure of words. Grew-Match can analyze the following Universals without needing any other tools:

- **Universal 6:** All languages with dominant VSO order have SVO as an alternative or as the only alternative basic order.
- **Universal 11:** Inversion of statement order so that verb precedes subject only occurs in languages where the question word or phrase is normally initial. This same inversion occurs in yes-no questions only if it also occurs in interrogative word questions.
- **Universal 12:** If a language has dominant order VSO in declarative sentences, it always puts interrogative words or phrases first in interrogative word questions; if it has dominant order SOV in declarative sentences, there is never such an invariant rule.
- **Universal 18:** When the descriptive adjective precedes the noun, the demonstrative, and the numeral, with overwhelmingly more than chance frequency, does likewise.
- **Universal 20:** When any or all of the items (demonstrative, numeral, and descriptive adjective) precede the noun, they are always found in that order. If they follow, the order is either the same or its exact opposite.
- **Universal 22:** If in comparisons of superiority, the only order, or one of the alternative orders, is standard-marker-adjective, then the language is postpositional. With overwhelmingly more than chance frequency if the only order is adjective-marker-standard, the language is prepositional.
- **Universal 31:** If either the subject or object noun agrees with the verb in gender, then the adjective always agrees with the noun in gender.

- **Universal 32:** Whenever the verb agrees with a nominal subject or nominal object in gender, it also agrees in number.
- **Universal 33:** When number agreement between the noun and verb is suspended and the rules is based on order, the case is always one in which the verb precedes and the verb is in the singular.
- **Universal 37:** A language never has more gender categories in non-singular numbers than in the singular.
- **Universal 40:** When the adjective follows the noun, the adjective expresses all the inflectional categories of the noun. In such cases the noun may lack over expression of one or all of these categories.
- **Universal 44:** If a language has gender distinctions in the first person, it always has gender distinctions in the second or third person, or in both.
- **Universal 45:** If there are any gender distinctions in the plural of the pronoun, there are some gender distinctions in the singular also.

20.3 Universals through UD Treebank Hub Page

The Universals that can be consulted on the UD Treebank Hub Page refer to morphosyntactic features. The general information of each corpus reveals whether the tags referring to the values stated in the Universal are activated or not. The following Universals can be checked directly from the Treebank Hub Page in Universal Dependencies (Nivre et al., 2023) (but also in Grew-Match):

- **Universal 30:** If the verb has categories of person-number or if it has categories of gender, it always has tense-mode categories.
- **Universal 34:** No language has a trial number unless it has a dual. No language has a dual unless it has a plural.
- **Universal 36:** If a language has the category of gender, it always has the category of number.
- **Universal 42:** All languages have pronominal categories involving at least three persons and two numbers.
- **Universal 43:** If a language has gender categories in the noun, it has gender categories in the pronoun.

20.4 Universals through Combination of Different Methods

Because of some elements contained in them, there are some Universals that require a combination of different methods to be tested. The reason for this is individual and different for each Universal and will be explained in detail in its development:

- **Universal 2:** In languages with prepositions, the genitive almost always follows the governing noun, while in languages with postpositions it almost always precedes.
- **Universal 5:** If a language has dominant SOV order and the genitive follows the governing noun, then the adjective likewise follows the noun.
- **Universal 7:** If in a language with dominant SOV order, there is no alternative basic order, or only OSV as the alternative, then all adverbial modifiers of the verb likewise precede the verb.
- **Universal 14:** In conditional statements, the conditional clause precedes the conclusion as the normal order in all languages.
- **Universal 23:** If in apposition the proper noun usually precedes the common noun, then the language is one in which the governing noun precedes its dependent genitive. With much better than chance frequency, if the common noun usually precedes the proper noun, the dependent genitive precedes its governing noun.
- **Universal 41:** If in a language the verb follows both the nominal subject and nominal object as the dominant order, the language almost always has a case system.

20.5 Universals with Current Limitations

There are some Universals that we cannot check with the analysis we propose. These refer to aspects of language that have not been considered in UD; therefore, we do not have this labeled information. We are referring either to intonational patterns or to morphological aspects (the structure of the words).

The following Universals cannot be checked nowadays with the current data in Universal Dependencies:

- **Universal 8:** When a yes-no question is differentiated from the corresponding assertion by an intonational pattern, the distinctive intonational features of each of these patterns are reckoned from the end of the sentence rather than from the beginning.
- **Universal 9:** With well more than chance frequency, when question particles or affixes are specified in position by reference to the sentence as a whole, if initial, such elements are found in prepositional languages, and, if final, in postpositional.
- **Universal 10:** Question particles or affixes, when specified in position by reference to a particular word in the sentence, almost always follow that word. Such particles do not occur in languages with dominant order VSO.
- **Universal 15:** In expressions of volition and purpose, a subordinate verbal form always follows the main verb as the normal order except in those languages in which the nominal object always precedes the verb.
- **Universal 26:** If a language has discontinuous affixes, it always has either prefixing or suffixing or both.
- **Universal 27:** If a language is exclusively suffixing, it is postpositional; if it is exclusively prefixing, it is prepositional.
- **Universal 28:** If both the derivation and inflection follow the root, or they both precede the root, the derivation is always between the root and the inflection.
- **Universal 29:** If a language has inflection, it always has derivation.
- **Universal 35:** There is no language in which the plural does not have some non-zero allomorphs, whereas there are languages in which the singular is expressed only by zero. The dual and the trial are almost never expressed only by zero.
- **Universal 38:** Where there is a case system, the only case which ever has only zero allomorphs is the one which includes among its meanings that of the subject of the intransitive verb.
- **Universal 39:** Where morphemes of both number and case are present and both follow or both precede the noun base, the expression of number almost always comes between the noun base and the expression of case.

In summary, the overview of the Universals we found can be summarised in the table 20.2.

Univer- sal	Typo- metrics	Grew	Hub Page	Combi- nation	Impossi- ble
1	X	X			
2		X		X	
3	X	X			
4	X	X			
5	X	X		X	
6		X			
7	X	X		X	
8					X
9					X
10					X
11		X			
12		X			
13	X	X			
14		X		X	
15					X
16	X	X			
17	X	X			
18		X			
19	X	X			
20		X			
21	X	X			
22		X			
23	X	X		X	
24	X	X			
25	X	X			
26					X
27					X
28					X
29					X
30		X	X		
31		X			
32		X			
33		X			
34		X	X		
35					X
36		X	X		
37		X			
38					X
39					X
40		X			
41	X	X		X	
42		X	X		
43		X	X		
44		X			
45		X			

Table 20.2: Overview of Universals and Tools

21

Formal Analysis of Universals

In this chapter, we will show the formalization we propose for each Universal, and we will explain the process we followed to reach this goal, organized by the different tools that can be used. In other words, we are going to offer a linguistic description that explains the formalization, and also we are going to justify the choices made. We explain in detail the steps we have taken so they can be helpful for future researchers. Then, we show the results we obtained in each of the Universals with the previously mentioned tools.

We believe that the results presented here are either novel or complementary to previous approaches and allow us to understand the original proposals of the American typologist in an updated way. The results shown below will be presented with the formalizations that have been discussed to facilitate reading. If it is possible to show a Typometrics graph, this will be done. We must remember that, for consistency with the whole work, the final objective is to present the results with the GLU model we have proposed.

However, we must also remember the necessary artifice for the functioning of the GLU model: the quantitative data. Although this exact data is irrelevant to the representation, it is relevant when configuring the model. Since we must have this data for the correct functioning of the model, we also decided to show it in case it interests other researchers who use an exclusively quantitative approach. This does not mean, therefore, that the data we show here contradicts the proposal of the thesis, which is different but is nourished by such data. We should also remember two other relevant aspects when representing the data.

1. Firstly, we show the result we obtained with Greenberg’s original proposal in the trichotomous models of each of the Universals. However, we also show, if possible, other Universal proposals in comparison with Greenberg’s original version, alternatives to it, or extensions to it. To show Greenberg’s original Universal, we will use the color green. To show our alternative proposals, we will use orange. If the Greenberg Universal is composed of different Universals, the general Universal (the definitive version) is shown in green. We show the small Universals within that Universal in blue. We reserve the yellow color to carry out some innovative technique that escapes the above description, although its development will be detailed in the given Universal.
2. Secondly, we should mention that once all the Universals have been presented individually, we carry out two more model proposals, which correspond to, on the one hand, a general and unified classification of all the original Greenberg Universals. On the other hand, a general and unified classification of the version of Greenberg’s Universals that we find most interesting is either Greenberg’s (marked with the color green) or some variant of ours (marked with a color other than green).

All the Universals that can be analyzed with UD can be checked via Grew-Match. In some of them, using a few extra steps or support tools is necessary. Some of these Universals can also be checked, thanks to Typometrics. Others can be checked directly on the UD Treebank Hub Page. That is, this information is easily checked (manually) in each treebank’s general information within the Universal Dependencies website (Nivre et al., 2023).

Then, we obtained the results. It should be remembered that we applied this generated knowledge to UD version 2.11, which was the most up-to-date at the time of analysis, and the SUD annotation scheme was selected, as we have already mentioned.

We should not forget that even today’s Linguistic Typology in computation is a growing reality but not the most widespread among researchers. We have even less research using Universal Dependencies, given the novelty of its existence and, especially, because we have only recently had versions with a sufficiently interesting variety of languages and corpora to begin to carry out studies of this type (although more languages still need to be added).

We do not observe many studies using the SUD annotation scheme, which implies, in our opinion, a better representation of the phenomena in Typology. In our view, the labels used are more accurate than those of UD. By way of example, in SUD no distinction is made between subject dependencies depending on whether this is produced in a nominal or clausal element.

As for the computational study of Greenberg’s proposals, we have not observed many studies beyond those already mentioned in this doctoral thesis. To our knowledge, there is no integral study of Greenberg’s original proposal; we have only seen isolated proposals commenting on some specific Universal and using different methodologies, as those we have previously criticized.

21.1 Universal 1

UNIVERSAL 1: *In declarative sentences with nominal subject and object, the dominant order is almost always one in which the subject precedes the object.*

21.1.1 How to Check It in Typometrics

Technical Decisions
1. Firstly, the Universal is analyzed with Typometrics (Gerdes et al., 2021). 2. We select the SUD annotation scheme because we can get results more comparable to the classical typological claims. 3. We select Graph 2D. 4. We set a filter of utterances to avoid misleading data. 5. We select <i>direction – cfc</i> in both x-Axis and y-Axis, meaning that we are interested in the word order feature but applied only in a specific Part of Speech. 6. In the x-Axis we select <i>VERB – subj – NOUN</i> and in the y-Axis we select <i>VERB – comp : obj – NOUN</i>. This measure is selected because, in the original Universal, it is formulated only for nominal objects and subjects, not pronominal, for example.

Table 21.1: Technical Decisions for Checking Universal 1 in Typometrics

We must mention that for this Universal, there have been some computational attempts from other researchers and methodologies recently (Choi, Guillaume, & Fort, 2021), so it would be interesting to compare results. We must warn that Typometrics does not allow filtering with or without interrogations and exclamations. Greenberg, on the other hand, did make this distinction and only took into account declarative sentences. So, in the case one wonders if these kinds of sentences can bias the results or if this effect can be neutralized (except for specific treebanks based exclusively on poetry, Q & A, etc.), we can compare the results obtained by Typometrics with the

results in Grew-Match (since in Grew-Match the sentence modality filter can be applied and we have done so).

Even if the tool has limitations, we want to offer a visual and easy-interpretable output (if possible) regarding the distribution of more subjects than objects before verbs. It is also interesting to check how pronouns behave (since they are typically excluded in this kind of characterization); for this, we should change both *direction – cfc*: the first one should be *VERB – subj – PRON*, and the second one should be *VERB – comp : obj – PRON* now.

21.1.2 Formalisation Proposed for Grew-Match

Even if we could directly ask at the same time for all the orders and patterns where the subject appears before the object, we preferred to perform separate queries regarding all the different possible word orders because that is what Greenberg originally stated (Greenberg et al., 1963). In the Universal, he says, “the dominant order is always one in which the subject precedes the object”. Then, it does not consider all possible orders at once but states that from all the possible orders, the dominant one will be one in which the subject is before the object, that is, SVO, SOV, or VSO. Then, we have to check all possible word orders, not only in a ranking position but to consider each word order’s occurrences. To determine it, we follow the criteria of the double, common in statistics in linguistics and already applied in other papers like Dryer (2013) or Choi (2021), well explained in Gries (2013:17), known as ratio variable in rank (ordinal variable). That is: “we consider the most frequent order as the dominant order for a given corpus provided that it is at least twice as frequent as the next most frequent” (Choi, 2021). In other words, this means that for each corpus, we observe the ratio between the number of occurrences of the most frequent order; if the ratio is greater than or equal to 2, the most frequent order is the dominant order else, we consider the corpus to exhibit a NDO (No Dominant Order).

For example, in Spanish (AnCora), we find 3952 occurrences of SVO, 39 of VOS, 26 of OVS, 26 of VSO, 4 of OSV, and 0 of SOV. Thus, we can confirm that for Universal 1, the dominant word order of Spanish is SVO, more than double the frequency compared to the second one. Maybe one can be surprised by the results obtained in Spanish, especially regarding 0 occurrences in SOV. However, we have to detail the formalization we have proposed for Greenberg’s Universal, which explains these data.

We have to remember that Greenberg is forming a Universal regarding declarative sentences and dominant word order. These two are the main restrictions of the Universal:

1. The fact of being a declarative sentence is already forcing us to exclude interrogations and exclamations that are not declarative (and that is what we do in the *without* pattern, negative filtering these symbols). Also, when discussing dominant word order in Typology, it is common to consider that only can be checked in as the word order of declarative sentences, considering the rest of them, marked productions (Mithun, 1992). For this, when basic word order is mentioned, we always have to avoid other constructions than the declarative one. One can think that indirect interrogative questions are still occurring with this filtering. However, we also have to remember that indirect questions are always placed in a declarative sentence, so we are dealing with a declarative sentence.
2. The second restriction found in Greenberg is regarding the part of speech of the subject and the direct object, as we mentioned. The linguist explicitly mentions “nominal subject and object”, the less marked structure regarding word order, which corresponds to the type of analysis of searching the neutral word order according to the different ways presented in Mithun (1992). Then, we must discard all subjects and objects expressed through a pronoun as the main dependent (but not those containing a pronoun that is just a modifier of this main dependent, for example). For example, these two restrictions cause us to find 0 evidence of SOV order in Spanish. It is important to note that in different moments of his paper, Greenberg insists that his typological claims “will be the relative order of subject, verb, and object in declarative sentences with nominal subject and object”. Thus, even if we think there can be other possibilities to capture the basic word order of languages, if we want to formalize Greenberg’s Universals and offer a usage-based approach to them, we have to follow his claims.

To summarize, to formalize Greenberg’s Universals, we must determine basic word order.

1. considering only declarative sentences;
2. the presence of verbs, objects, and subjects;
3. and the last two have to be a noun (common or proper).

Debating other alternatives to capture word order in languages that may offer new and more detailed results is interesting. Thus, even if we are constrained to formalize Greenberg Universals by the methodology described in the previous paragraph, we want to propose in Universal 1, (section 21.1) different possible word order formalizations that could also show valuable results. We will

formalize some of these possibilities and show how results can change in considering Universal 1. This would even lead to a new possible typological characterization of word order in languages regarding specific contexts or with specific restrictions. However, we prefer to let this objective be a future work born in this dissertation since this would go beyond what we are proposing in our study, and we would lose the focus. However, as we said, we want to open the path for future work by setting different alternatives, as we will see.

Regarding the patterns created for the formalization of the 1st Greenberg Universal, we find, first of all, the information relative to the pattern created. After the “%” symbol, as stated in the symbol tables to be used for extra info or comments, it is placed the number of the Universal and the linguistic pattern we are formalizing. This is done in all created patterns.

We propose the following patterns (21.1) - (21.6).

The same elements are formalized in all six different patterns. We ask for a word that is a verb to be the head of a noun or proper noun in a subject dependency relation and this verb to be also the head of a noun or proper noun in a direct object dependency relation. In all six patterns, we also apply the same restriction (*without* pattern): we do not want to count any utterances that have a punctuation symbol “?” or “!” depending on any verb of the sentence (since these symbols always depend on a verb).

The difference in the six patterns formalized is the restriction regarding the order of these elements.

$$\begin{array}{l}
 \%1 - SVO \\
 pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = \\
 PROPEN|NOUN]; V - [comp : obj] - > O; O[upos = \\
 PROPEN|NOUN]; V >> S; S << O; V << O\} \\
 without\{V - [punct] - > P; P[lemma = “?”|“!”]\}
 \end{array} \quad (21.1)$$

To capture the SVO word order (21.1), we ask the subject to be before the verb and the object to be after the verb by using different order restrictions in these three elements with “<<” or “>>”, as shown in the symbol table.

$$\begin{array}{l}
 \%1 - SOV \\
 pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = \\
 PROPEN|NOUN]; V - [comp : obj] - > O; O[upos = \\
 PROPEN|NOUN]; V >> S; S << O; V >> O\} \\
 without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
 \end{array} \quad (21.2)$$

To capture the SOV word order (21.2), we ask the subject to be before the verb and the object to be before the verb but not before the subject.

$$\begin{array}{l}
 \%1 - VSO \\
 pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = \\
 PROPEN|NOUN]; V - [comp : obj] - > O; O[upos = \\
 PROPEN|NOUN]; V << S; S << O\} \\
 without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
 \end{array} \quad (21.3)$$

In VSO case (21.3), we want a verb before the subject and before the object, but the subject before the object.

$$\begin{array}{l}
 \%1 - VOS \\
 pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = \\
 PROPEN|NOUN]; V - [comp : obj] - > O; O[upos = \\
 PROPEN|NOUN]; V << S; S >> O; V << O\} \\
 without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
 \end{array} \quad (21.4)$$

In VOS word order (21.4), we formalize the restriction of the verb being before the subject and before the object, but the object appearing before the subject.

$$\begin{array}{l}
 \%1 - OSV \\
 pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = \\
 PROPEN|NOUN]; V - [comp : obj] - > O; O[upos = \\
 PROPEN|NOUN]; V >> S; S >> O; V >> O\} \\
 without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
 \end{array} \quad (21.5)$$

In the OSV order (21.5), we ask for objects appearing before subjects and verbs, but subjects before verbs.

$$\begin{array}{l}
\%1 - OVS \\
pattern\{V[upos = VERB]; V - [subj]- > S; S[upos = \\
PROPN|NOUN]; V - [comp : obj]- > O; O[upos = \\
PROPN|NOUN]; V << S; S >> O; V >> O\} \\
without\{V - [punct]- > P; P[lemma = "?"|"!"]\}
\end{array} \quad (21.6)$$

In OVS order (21.6), we ask for objects appearing before subjects and verbs, but verbs before subjects.

We can interpret the data once we have checked these patterns in all corpora. Nevertheless, to sum up, the correct interpretation of the results involves necessarily these previous steps:

1. Ranking each order in each language from the most to the least frequent.
2. Checking the occurrences in each order.
3. If the first one is double (or more) than the second one, this language is considered as having the dominant order of this first order. If not, this language is considered a nondominant order language.
4. If the dominant order is SVO, SOV, or VSO, the language is counted as an example that contributes positively to the Universal.
5. If the dominant order is VOS, OSV, or OVS, the language is considered an exception or counterexample of the formulated Universal.

If the final result is “nondominant order”, there are three possible interpretations, given the ambiguity in Greenberg’s formulation and his lack of clarity with these cases.

1. On the one hand, it can be understood that if a language has “NDO”, it cannot be considered a language with a dominant order where the subject comes before the object, so it must be counted as an exception.
2. Another possibility is to consider NDOs as languages that escape the classification proposed by Greenberg and, therefore, should be discarded from the study.
3. Finally, we can also see if this NDO is caused by a two-order competition where the subject comes before the object (SVO and SOV, for example). In that case, there would also be the possibility of counting such a language NDO as one in which the Universal is fulfilled.

As can be seen, none of these interpretations of how to treat NDO languages is right or wrong and have not been detailed previously. To be more rigorous with the Universals proposed by Greenberg, we believe we should be as restrictive and detailed as possible, as happens with the first scenario.

If we would like to check, in general, the behavior of pronouns in Grew-Match, we should apply the following formalizations (21.7) - (21.12):

$$\begin{array}{l}
 \%1 - SVO_{pn} \\
 pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = PRON]; V - \\
 [comp : obj] - > O; O[upos = PRON]; V >> S; S << O; V << O\} \\
 without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
 \end{array} \tag{21.7}$$

In (21.7), we check SVO basic word order where the nominal elements are pronouns.

$$\begin{array}{l}
 \%1 - SOV_{pn} \\
 pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = PRON]; V - \\
 [comp : obj] - > O; O[upos = PRON]; V >> S; S << O; V >> O\} \\
 without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
 \end{array} \tag{21.8}$$

In (21.8), we check SOV basic word order where the nominal elements are pronouns.

$$\begin{array}{l}
 \%1 - VSO_{pn} \\
 pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = PRON]; V - \\
 [comp : obj] - > O; O[upos = PRON]; V << S; S << O\} \\
 without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
 \end{array} \tag{21.9}$$

In (21.9), we check VSO basic word order where the nominal elements are pronouns.

$$\begin{array}{l}
 \%1 - VOS_{pn} \\
 pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = PRON]; V - \\
 [comp : obj] - > O; O[upos = PRON]; V << S; S >> O; V << O\} \\
 without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
 \end{array} \tag{21.10}$$

In (21.10), we check VOS basic word order where the nominal elements are pronouns.

$$\begin{array}{l}
\%1 - OSVpn \\
pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = PRON]; V - \\
[comp : obj] - > O; O[upos = PRON]; V >> S; S >> O; V >> O\} \\
without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
\end{array} \tag{21.11}$$

In (21.11), we check OSV basic word order where the nominal elements are pronouns.

$$\begin{array}{l}
\%1 - OVS pn \\
pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = PRON]; V - \\
[comp : obj] - > O; O[upos = PRON]; V << S; S >> O; V >> O\} \\
without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
\end{array} \tag{21.12}$$

In (21.12), we check OVS basic word order where the nominal elements are pronouns.

The formalized patterns we offer behave the same way as the ones made in (21.1) - (21.6). The only change we applied is that the part of speech is no longer *NOUN* or *PROPN* (noun or proper noun) but *PRON* (pronoun). The word order restrictions are maintained in each of the six different word orders by the orders of SVO, SOV, VSO, VOS, OSV, and OVS shown before through the “<<” restriction.

We must warn that this is just a general approach to the complex phenomenon of pronouns. The results shown can draw a tendency of the behavior of pronouns, but there are further filters that would enrich the study of pronoun word order regarding different constraints.

Another proposal we make is to consider both of the formalizations presented before. That is, to check the word order of subjects and objects that are common nouns, proper nouns, or pronouns. To create this pattern, we should reuse the first one proposed in (21.1) - (21.6), but add the *PRON* value in the possible *upos*. That is, we would have the patterns (21.13) -(21.18):

$$\begin{array}{l}
\%1 - SVO pnN \\
pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = \\
PROPN|NOUN|PRON]; V - [comp : obj] - > O; O[upos = \\
PROPN|NOUN|PRON]; V >> S; S << O; V << O\} \\
without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
\end{array} \tag{21.13}$$

Specifically, in (21.13), we characterize basic SVO order whose subject and direct object can be nominal or pronominal.

$$\begin{array}{l}
\%1 - SOVpnN \\
pattern\{V[upos = VERB]; V - [subj]- > S; S[upos = \\
PROPN|NOUN|PRON]; V - [comp : obj]- > O; O[upos = \\
PROPN|NOUN|PRON]; V >> S; S << O; V >> O\} \\
without\{V - [punct]- > P; P[lemma = "?"|"!"]\}
\end{array} \tag{21.14}$$

Specifically, in (21.14), we characterize basic SOV order whose subject and direct object can be nominal or pronominal.

$$\begin{array}{l}
\%1 - VSOpnN \\
pattern\{V[upos = VERB]; V - [subj]- > S; S[upos = \\
PROPN|NOUN|PRON]; V - [comp : obj]- > O; O[upos = \\
PROPN|NOUN|PRON]; V << S; S << O\} \\
without\{V - [punct]- > P; P[lemma = "?"|"!"]\}
\end{array} \tag{21.15}$$

Specifically, in (21.15), we characterize basic VSO order whose subject and direct object can be nominal or pronominal.

$$\begin{array}{l}
\%1 - VOSpnN \\
pattern\{V[upos = VERB]; V - [subj]- > S; S[upos = \\
PROPN|NOUN|PRON]; V - [comp : obj]- > O; O[upos = \\
PROPN|NOUN|PRON]; V << S; S >> O; V << O\} \\
without\{V - [punct]- > P; P[lemma = "?"|"!"]\}
\end{array} \tag{21.16}$$

Specifically, in (21.16), we characterize basic VOS order whose subject and direct object can be nominal or pronominal.

$$\begin{array}{l}
\%1 - OSVpnN \\
pattern\{V[upos = VERB]; V - [subj]- > S; S[upos = \\
PROPN|NOUN|PRON]; V - [comp : obj]- > O; O[upos = \\
PROPN|NOUN|PRON]; V >> S; S >> O; V >> O\} \\
without\{V - [punct]- > P; P[lemma = "?"|"!"]\}
\end{array} \tag{21.17}$$

Specifically, in (21.17), we characterize basic OSV order whose subject and direct object can be nominal or pronominal.

$$\begin{array}{l}
\%1 - OVS_{pnN} \\
pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = \\
PROP_N|NOUN|PRON]; V - [comp : obj] - > O; O[upos = \\
PROP_N|NOUN|PRON]; V << S; S >> O; V >> O\} \\
without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
\end{array} \tag{21.18}$$

Specifically, in (21.18), we characterize basic OVS order whose subject and direct object can be nominal or pronominal.

Another interesting possibility is to change the methodology of reaching the basic word order of a language. That is, according to the different theories presented by Mithun (1992), it would be to change the mix of frequency in word order with a neutral word order restriction, the one proposed by Greenberg (1963), to one that is purely based on frequency. That would mean we should consider any possible part of speech that could potentially depend on the subject or object dependency. This proposal of formalization is quite interesting because, thanks to computational data based on a corpus, we can reach all the less frequent and borderline cases that are difficult to find in grammars, which tend to offer only the most prototypical cases, or, for example, is what we find in Greenberg (1963), since when prepositions (adpositions before) are mentioned, only their relation with nouns is taken into account.

Then, to formalize it, we have to copy the structure created for the first patterns (21.1) - (21.6), but we have to formalize that potentially any part of speech can act as a subject or object. Thus, we have to select all the part of speech tags available. However, we decide to exclude two of them: *PUNCT* and *X*, since these two tags cannot offer occurrences because a punctuation symbol (like a dot “.”) always receives the dependency of *punct*, and what it is tagged under the *X* label are just noises or similar elements. One can think that it is difficult to see a particle (*PART*), for example, being the subject of a sentence and that maybe should not be considered in the formalization. We propose considering all the parts of speech that could potentially act as a subject. First of all, we cannot assure 100% that in any language of the world, a particle, following the example, could act as a subject. Secondly, our aim is also to propose a formalization (that we will bring to the practical and computational testing, too) that theoretically represents what a subject or object is. In this specific view, we are proposing, purely quantitatively, that we want to capture all subjects and objects. Thus, since the formalisation includes all the categories that could appear (but not necessarily have to if the language hasn’t it), we believe this is the best proposal in Universal terms. The parts of speech present in the pattern are mere possibilities, but if a language does not

have all of them (or not all of them for this dependency), they are just not counted.

If we are planning to capture quantitatively any utterance in any language that has subjects or objects without any neutrality restriction, we have to leave the restriction regarding interrogative and exclamation symbols since we are not asking only for declarative sentences anymore (since declarative sentences are being asked in the neutral sentence type word order research). We have to mention that *AUX* is not considered as a possible part of speech to be added into *V* category (that is, acting as a head of a subject or object) because an *AUX* form is always going to be a dependent of a *VERB* head in a *comp : aux* dependency relation (or we would have attributes and not objects, for example) and, then, *VERB* is always going to have the dependents *subj* or *comp : obj*, never *AUX*. The patterns we propose are (21.19) - (21.24):

$$\begin{array}{l}
 \%1 - SVO_{any} \\
 pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = \\
 ADJ|ADV|INTJ|NOUN|PROPN|VERB|ADP|AUX| \\
 CCONJ|DET|NUM|PART|PRON|SCONJ|SYM]; V - \\
 [comp : obj] - > O; O[upos = \\
 ADJ|ADV|INTJ|NOUN|PROPN|VERB|ADP|AUX| \\
 CCONJ|DET|NUM|PART|PRON|SCONJ|SYM]; V >> S; S << \\
 O; V << O\}
 \end{array} \tag{21.19}$$

In (21.19), we find basic SVO order with subjects and objects of any part of speech.

$$\begin{array}{l}
 \%1 - SOV_{any} \\
 pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = \\
 ADJ|ADV|INTJ|NOUN|PROPN|VERB|ADP|AUX| \\
 CCONJ|DET|NUM|PART|PRON|SCONJ|SYM]; V - \\
 [comp : obj] - > O; O[upos = \\
 ADJ|ADV|INTJ|NOUN|PROPN|VERB|ADP|AUX| \\
 CCONJ|DET|NUM|PART|PRON|SCONJ|SYM]; V >> S; S << \\
 O; V >> O\}
 \end{array} \tag{21.20}$$

In (21.20), we find basic SOV order with subjects and objects of any part of speech.

$$\begin{array}{l}
\%1 - VSOany \\
pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = \\
ADJ|ADV|INTJ|NOUN|PROPN|VERB|ADP|AUX| \\
CCONJ|DET|NUM|PART|PRON|SCONJ|SYM]; V - \\
[comp : obj] - > O; O[upos = \\
ADJ|ADV|INTJ|NOUN|PROPN|VERB|ADP|AUX| \\
CCONJ|DET|NUM|PART|PRON|SCONJ|SYM]; V << S; S << \\
O\}
\end{array} \tag{21.21}$$

In (21.21), we find basic VSO order with subjects and objects of any part of speech.

$$\begin{array}{l}
\%1 - VOSany \\
pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = \\
ADJ|ADV|INTJ|NOUN|PROPN|VERB|ADP|AUX| \\
CCONJ|DET|NUM|PART|PRON|SCONJ|SYM]; V - \\
[comp : obj] - > O; O[upos = \\
ADJ|ADV|INTJ|NOUN|PROPN|VERB|ADP|AUX| \\
CCONJ|DET|NUM|PART|PRON|SCONJ|SYM]; V << S; S >> \\
O; V << O\}
\end{array} \tag{21.22}$$

In (21.22), we find basic VSO order with subjects and objects of any part of speech.

$$\begin{array}{l}
\%1 - OSVany \\
pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = \\
ADJ|ADV|INTJ|NOUN|PROPN|VERB|ADP|AUX| \\
CCONJ|DET|NUM|PART|PRON|SCONJ|SYM]; V - \\
[comp : obj] - > O; O[upos = \\
ADJ|ADV|INTJ|NOUN|PROPN|VERB|ADP|AUX| \\
CCONJ|DET|NUM|PART|PRON|SCONJ|SYM]; V >> S; S >> \\
O; V >> O\}
\end{array} \tag{21.23}$$

In (21.23), we find basic OSV order with subjects and objects of any part of speech.

$$\begin{array}{l}
 \%1 - OVSany \\
 pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = \\
 ADJ|ADV|INTJ|NOUN|PROPN|VERB|ADP|AUX| \\
 CCONJ|DET|NUM|PART|PRON|SCONJ|SYM]; V \\
 [comp : obj] - > O; O[upos = \\
 ADJ|ADV|INTJ|NOUN|PROPN|VERB|ADP|AUX| \\
 CCONJ|DET|NUM|PART|PRON|SCONJ|SYM]; V << S; S >> \\
 O; V >> O\}
 \end{array} \tag{21.24}$$

In (21.24), we find basic OVS order with subjects and objects of any part of speech.

Another possibility that is worth formalizing, in order to compare with the results obtained by the main patterns, is to follow what some researchers after Greenberg have stated (Dryer, 1992, 1997, 2013; Lehmann, 1978; Vennemann, 1974, 1976): that even if they do believe that the relation between subjects, objects and verbs is the most powerful predictor in Typology, claims are more precise when being done separately. They would like to check SV and VS orders, then OV and VO orders. The main differences in this approach are two.

1. First, we would obtain more utterances because the pattern is less restrictive. We would obtain sentences with the three elements present and sentences with subject and verb but not object (e.g., intransitive verbs). The same thing would happen when checking verbs and objects.
2. Second, the previously cited researchers, especially Dryer (2013), believe that the most powerful distinction is between OV and VO languages, putting the SV-VS distinction in a second level.

Thus, we propose (21.25) - (21.28):

$$\begin{array}{l}
 \%1 - SV \\
 pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = \\
 PROPN|NOUN]; V >> S\} \\
 without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
 \end{array} \tag{21.25}$$

(21.25) corresponds to occurrences of a subject before a verb.

$$\begin{array}{l}
\%1 - VS \\
pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = \\
PROP N|NOUN]; V << S\} \\
without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
\end{array} \tag{21.26}$$

(21.26) corresponds to verb occurrences before a subject.

$$\begin{array}{l}
\%1 - OV \\
pattern\{V[upos = VERB]; V - [subj] - > O; O[upos = \\
PROP N|NOUN]; V >> O\} \\
without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
\end{array} \tag{21.27}$$

(21.27) corresponds to occurrences of an object before a verb.

$$\begin{array}{l}
\%1 - VO \\
pattern\{V[upos = VERB]; V - [subj] - > O; O[upos = \\
PROP N|NOUN]; V << O\} \\
without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
\end{array} \tag{21.28}$$

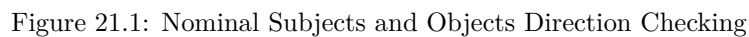
(21.28) corresponds to occurrences of a verb before an object.

In this case, we must simply rescue the first patterns again and eliminate, in the case of the 21.25 and 21.26, everything related to the direct object ($O[upos = NOUN|PROP N]$, $V - [comp : obj] - > O \ \& \ V << O$). In the case of 21.27 AND 21.28, we must keep everything from the original patterns, except delete ($S[upos = NOUN|PROP N]$, $V - [subj] - > S \ \& \ V << S$).

We can also apply the rest of the patterns previously developed for all three constituents (S, V & O), this time on only 2 (S & V or O & V), which would be interesting (pronouns only, pronouns and nouns, or any part of speech). However, we do not develop the repetition of these patterns this time in 2 constituents so as not to be repetitive. All the formalizations we have used are in the paper's appendices in case the reader is interested in consulting or using them.

21.1.3 Graphs from Typometrics

The first graph comparable to Greenberg's Universal using Typometrics is shown in figure 21.1.



We do not see any language in this part of the graph. Another aspect observed in this graph is the general distribution of all the languages, marked with the blue triangle in figure 21.2.

In this case, it should be noted that all languages are placed in the triangle at the top left, leaving the opposite one free. The few languages that can technically appear in such a triangle appear right on the border between these two triangles. The fact that they are all located in such a position and that we are working with specific corpora that can slightly modify the position of languages means that such residual cases should not be considered. In general, we can see how

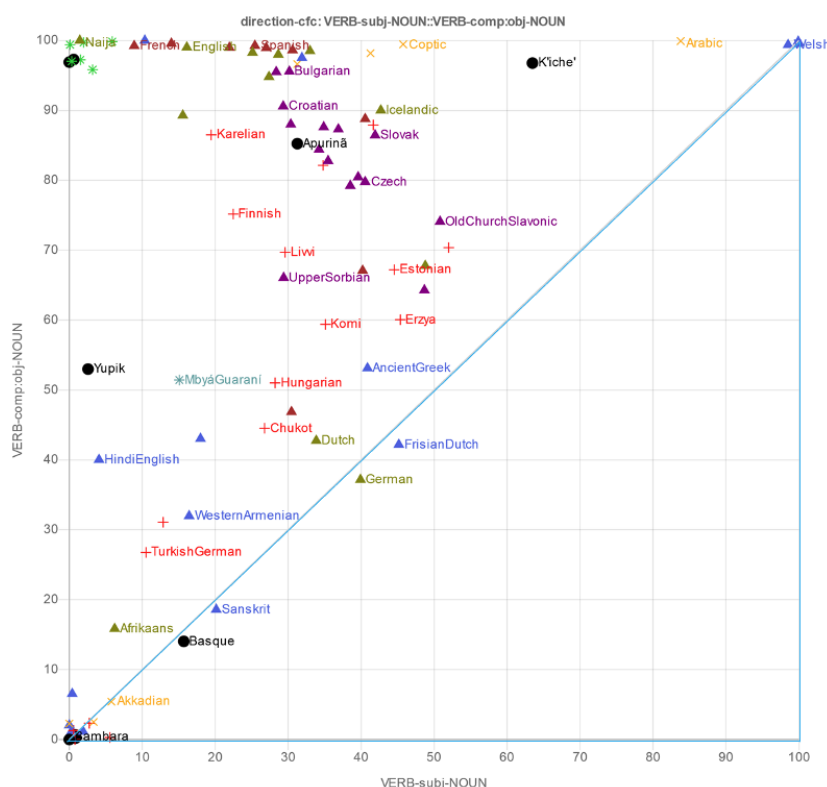


Figure 21.2: Nominal Subjects and Objects Distribution Proposal

languages, in other words, always have an equal or higher proportion of post-verbal direct object complements than of post-verbal subjects. This is, in short, further evidence to argue that there is a preference for subjects over direct objects (in this case, nominal) in such cases of precedence.

We can also check what happens with pronouns, as shown in figures 21.3 and 21.4.

In this case, some interesting differences can be observed.

1. First of all, the languages have quite a different behavior. In particular, what is observed is a general shift of the cluster that was more or less between 20 and 40 on the x-axis and 100 and 80 on the y-axis, which is now at 0-20 on the x-axis and 10-40 on the y-axis, more or less. This shows that languages that tend to be rather extreme with the postposition of the nominal object tend to lean towards an anteposed pronominalisation. We also observe that those same languages, which tend to place the nominal subject before the verb in most cases, place the pronominal subject before the verb as well, but in a more extreme way (check Roman Languages, Slavic Languages, and Finno-Ugric ones). In the case of nominally extreme languages with the precedence of subjects and objects, they do not vary in the shift to the pronoun version. In short, pronominalisation is usually a trigger towards the anteposition of

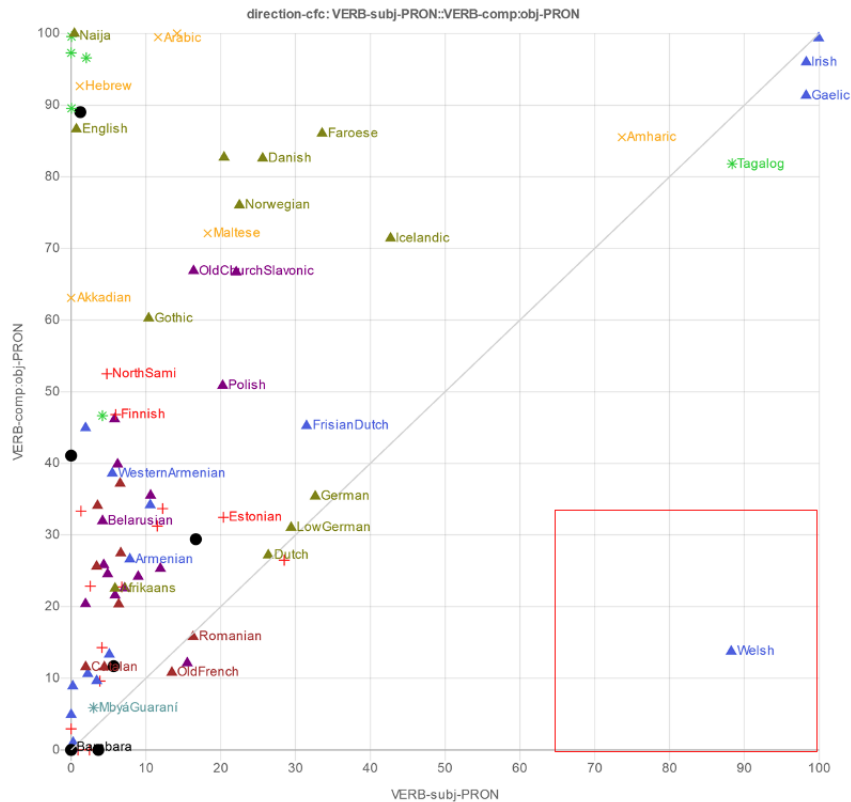


Figure 21.3: Pronominal Subjects and Objects Distribution Checking

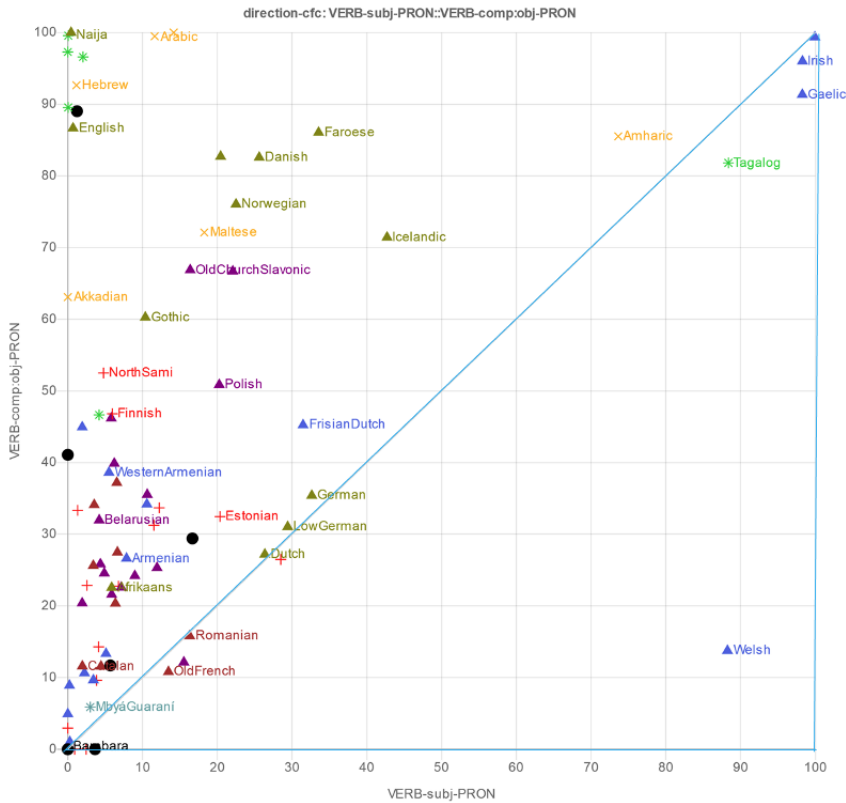


Figure 21.4: Pronominal Subjects and Objects Distribution Proposal

subjects and objects, whatever the configuration of the language. This falls squarely within the field of action of Universal 25, so we leave its expansion for its corresponding section.

2. Secondly, we see how a language now does appear in the red square where no languages were expected according to the Universal (Welsh) in the figure 21.3. Therefore, it seems to be a less predictive formulation. Furthermore, if we look at the figure 21.4, we can see how, although the languages that lie on the boundary between the two triangles are still in that area, there is an addition to the triangle that we expect to find empty at the top. Irish, Gaelic, and Tagalog (apart from the aforementioned Welsh) fall into this triangle, which gives us valuable information: in pronominal structures, VSO languages tend to show different behavior. In this case, this behavior refers to a shift from having the same proportion of subjects and objects postponed to the verb (the maximum possible) to showing some occurrences before, especially of direct objects, when analyzing pronouns (and in a very extreme way in the case of Welsh).

21.1.4 Trichotomous Expression Based on Grew-Match Data

Once we have seen the distribution in Typometrics, we can present the results obtained via Grew-Match, which gives us the exact quantitative data, perhaps useful for typologists and computational linguists and, in our case, an essential part in order to be able to model with the evaluative expression we use.

1. The original Universal proposed by Greenberg, strictly nominal, as we have reflected in the formalizations titled, is satisfied in 90% of the languages in the set (and is not satisfied in 10%).
2. In the case of the same check, changing nominal elements for pronominal ones, we obtain 72% compliance.
3. Adding occurrences of nominal and pronominal structures, we find compliance in 78% of the cases.
4. Without establishing any parts of speech restriction, we obtained 81% compliance with the Universal.

If we want to convert these results into the trichotomous model, we propose to understand Universals in language better. We obtain the result shown in figure 21.5.

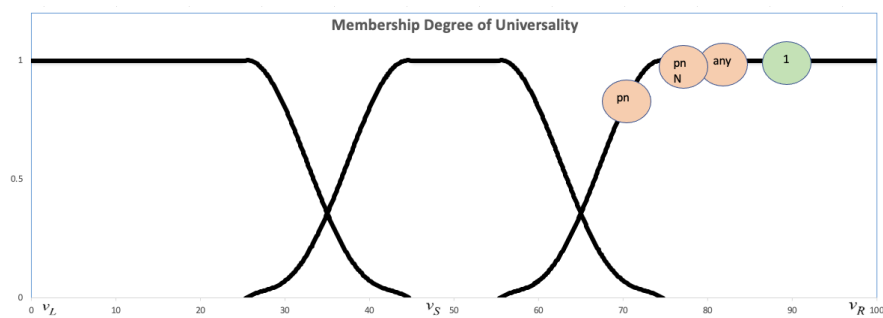


Figure 21.5: Trichotomous Expression for Universal 1

In this case, we can see how the Universal proposed by Greenberg is a High-Universal. Moreover, it is prototypical of its group, i.e., it is undoubtedly a clear example of such a Universal. The other formalization alternatives (using pronouns, pronouns, and nouns or without restricting parts of speech) obtain similar or worse results. In any case, we should mention that we are always talking about High-Universals, but we have been able to verify objectively that the choice made by Greenberg is not random or imprecise at all. Therefore, these results invite us to believe with greater certainty that Greenberg's proposal is entirely correct and that the terms in which it has been written are the ones that provide the best representation of the world's languages and their behavior.

In short, we must take for granted the tendency of the world's languages to place the nominal subject before the nominal object in unmarked sentences. It is also relevant to note that, although the data available in UD are biased in representing the orders that are more coincident with this Universal (especially SVO), the fact of presenting the result as a mere High-Universal allows us to approach such a restriction more respectfully. On the one hand, when we add languages or corpora in the future, the green sphere may shift slightly to the right, to the left, or remain unchanged, but in any case, it will still represent a High-Universal. On the other hand, this lack of languages present in a probability sample of under-represented macro-areas like Australia, for example, is compensated by this approach not based on exact quantities, which can also show the trend.

If the purpose of this thesis were not a direct contrast to Greenberg's formulation, a purely quantitative investigation could also be carried out, as proposed by researchers in the field of Typometrics. In this case, we should dispense with general categories and avoid dominance. In such a case, we should simply observe the results in which there is a higher proportion of one element than another. In other words, the results would be the ones in table 21.2.

Formalization	Accomplishment
Q-Original	100%
Q-Pron	95%
Q-Pron+N	98%
Q-AnyPoS	98%

Table 21.2: Accomplishment of Different Formalizations of Universal 1 in a Quantitative-Typometrical Way

In summary, we can see that in 100% of the languages analyzed, there are more orders with the subject preceding the direct object. However, this is not the dominant order in the case of the formulation in Grenberg’s terms (nominal elements). In the case of pronominal elements, we found 95% compliance, and in the case of the sum of pronouns and nouns or without part of speech restriction, we found 98% compliance.

As can be seen by this approach, the results are pretty interesting, which reinforces Greenberg’s proposal insofar as, by other parameters, such a restriction is still fulfilled. However, it is impossible to compare both results directly since the analysis methodology is different. That is, an analysis in purely quantitative terms (not in terms of dominance) would imply not a revision of Greenberg’s Universals but a new proposal of these through reformulating his study methodology. For example, it does not allow working with the nondominant order category, which is essential for the Typology in Greenberg. However, once we have reviewed Greenberg’s Universals in this thesis, we propose as an avenue for future work the proposal of such Universals using techniques other than those proposed in this study and with different metrics. We can summarise our findings as in table 21.3.

Summary of characteristics of the Universal 1 analysis:
1. Typometrics: restriction on VS + OV languages (free lower right corner).
2. Typometrics: a higher proportion of objects after the verb than subjects (free lower triangle).
3. Typometrics: similar behavior with pronouns (except Welsh).
4. Fulfilment of 90% of the original Universal.
5. The original Universal is a High-Universal.
6. Less strong (but still High-Universal) results in parallel checks: no restriction on the part of speech, pronouns, pronouns + nouns.

Table 21.3: Summary of Characteristics of the Universal 1 Analysis

21.2 Universal 2

UNIVERSAL 2: *In languages with prepositions, the genitive almost always follows the governing noun, while in languages with postpositions, it almost always precedes.*

21.2.1 How to Partially Check the Lemma Clustering

First, the Universal asks us to determine whether the language belongs to the prepositional or postpositional type, the first part of the implication. This check is straightforward and has already been dealt with in other Universals in this thesis. We must propose something like in the patterns (21.49) and (21.50). We recall that in these cases, we ask for the occurrence of a noun and an adposition since Greenberg’s characterization is based on the canonical dependence of the adposition, i.e., with a noun. We establish a *comp* : *obj* dependency relation, since it is an obligatory or argument dependency, between these two elements, in which the adposition is the head and the name a dependent. To capture prepositions, we require the adposition to precede the name ($ADP << N$) and to capture postpositions; we require the name to precede the adposition ($N << ADP$).

However, in these cases, we can also review less canonical relations between the adposition and other dependent parts of speech, as we have proposed in Universal 3 (21.3), for example. In this case, we must rescue the formalizations already proposed corresponding to (21.51), (21.52), (21.53), (21.54), (21.55), (21.56), (21.57) and (21.58). We must remember that in the case of (21.55) and (21.56), we are looking for the opposite sense of occurrence since, in this case, the adposition ceases to be the head and becomes the dependent of a verb, as we have already seen.

Regarding the second part of the Universal, the position of the genitive, we must say that its formalization and search in the UD tool is much more complex since languages do not have a single way of representing this concept. Therefore, we did not find any tag referring to the specific genitive relation.

In the situation where the language expresses this relation using a specific case (genitive case), as Polish, for example, it is easily verifiable. We can propose the formalizations (21.29) and (21.30) to determine whether the order of the words with a genitive case is before or after the head of this dependency.

$$\begin{array}{l}
\%2 - Gen - aft \\
pattern\{N[upos = NOUN]; GEN[Case = Gen]; N - [mod] - > \\
GEN; N << GEN\}
\end{array} \tag{21.29}$$

To capture how the dependent with genitive case appears after the nominal head.

$$\begin{array}{l}
\%2 - Gen - bef \\
pattern\{N[upos = NOUN]; GEN[Case = Gen]; N - [mod] - > \\
GEN; N >> GEN\}
\end{array} \tag{21.30}$$

To capture how the dependent with genitive case appears before the nominal head.

However, many languages express the genitive using a specific adposition, such as the preposition “di” in Italian. In these cases, we do not know which structure induces the genitive in each language. Since we do not have the information in interglossae either, one option would be to obtain that lexical information in another tool, similar to what we propose in Universal 14 in PanLex. Another option is to filter the adpositions of the corpora and keep the most frequent adposition in each language. This can be done by searching for adpositions ($patternADP[upos = ADP]$) and then adding a “key” to create clusters from the given restriction and add “ $ADP.lemma$ ”, which will allow us to group the total occurrences by lemmas. Presumably, this most frequent adposition is the one that functions as a genitive, the other uses being much less frequent. For example, continuing with Italian, in the ISDT corpus, we find precisely “di” as the most frequent preposition, with 19512 occurrences. The second most frequent adposition would be “a”, with 7661 occurrences.

Other languages, such as English, present the two linguistic types described above simultaneously, since on the one hand, we have the preposition “of” and, on the other hand, we have the Saxon genitive (‘s). Moreover, there is the difficulty of the treatment of the Saxon genitive since there are annotators who may assimilate it to other genitive case marks (such as “-i” or “-y” for singular masculines in Polish) and therefore mark it directly as a feature of the noun and others may understand it as a particle (PART). In fact, in the UD environment, there is often this tendency to consider it a particle.

Of course, we must not forget our little knowledge of agglutinative languages in Universal Dependencies, which is a clear limitation for studying structures of this type.

On the other hand, we have languages, such as Vietnamese, which have no genitive case marking but also use no adposition. In these cases, we find two contiguous nouns, one equivalent to a genitive

structure. This is what we believe most closely resembles the behavior of a Creole language. On the other hand, languages such as Thai have an adposition used for the genitive, but this is not obligatory. That is to say, in many cases, this adposition falls, and we find a structure very similar to that of Vietnamese.

Finally, we must not forget other cases of mixtures of different tools, such as Catalan, for example, where the main use of the genitive is through the preposition “de”. However, we also find the expression of genitive in pronouns. Although this case is similar to what we find in English, in languages such as Catalan, we find the particularity that the use of one or the other is restricted to the part of speech in question, and in this case, there is no possibility of expressing the structure through the nomenclature of a particle, but it is added as a feature to the pronoun.

In short, we do not have the facilities to ensure with certainty the structures used by each of the world’s languages (at least in a computational way without requiring grammars). Moreover, the mere fact of solving the treatment of genitives in the languages of the world in a computational way could give rise to an independent thesis, given the complexity of the case. Therefore, we prefer to focus on the ascertainment of genitive cases, in a restricted sense of Greenberg’s Universal, in order to be able to propose an initial trend that can be later confirmed, extended, or contradicted when more languages and linguistic types are added. Therefore, we leave the definitive resolution of the treatment of the genitive in the languages of the world as a new door that we open for future work.

It should be noted, moreover, that the brief overview we have given of the manifestations of genitive is not exhaustive but simply attempts to show a general panorama of the situation in the languages of the world in order to highlight the problem we encounter when investigating it from a Universal point of view.

21.2.2 Formalization Proposed for Grew-Match

Without serving as a definitive formalization, since, as we have commented, we believe that it is a rather deep subject, we can formalize cases that help us investigate the complex subject of the genitive. In this case, if we want to try to trace the cases in which it manifests itself through an adposition, we propose the possibilities (21.31) - (21.36):

$$\begin{array}{l}
\%2 - Gen - aft - adp \\
pattern\{N[upos = NOUN]; ADP[upos = ADP]; N - [mod] - > \\
ADP; N << ADP\}
\end{array} \tag{21.31}$$

We use (21.31) to capture how a modifying genitive by adposition appears after the nominal head.

$$\begin{array}{l}
\%2 - Gen - bef - adp \\
pattern\{N[upos = NOUN]; ADP[upos = ADP]; N - [mod] - > \\
ADP; N >> ADP\}
\end{array} \tag{21.32}$$

We use (21.32) to capture how a modifying genitive by adposition appears before the nominal head.

$$\begin{array}{l}
\%2 - Gen - aft - adp - udep \\
pattern\{N[upos = NOUN]; ADP[upos = ADP]; N - [udep] - > \\
ADP; N << ADP\}
\end{array} \tag{21.33}$$

We use (21.33) to capture how a genitive by adposition appears after the nominal head.

$$\begin{array}{l}
\%2 - Gen - bef - adp - udep \\
pattern\{N[upos = NOUN]; ADP[upos = ADP]; N - [udep] - > \\
ADP; N >> ADP\}
\end{array} \tag{21.34}$$

We use (21.34) to capture how a genitive by adposition appears before the nominal head.

$$\begin{array}{l}
\%2 - Gen - aft - adp - any \\
pattern\{N[upos = NOUN]; ADP[upos = ADP]; N - > ADP; N << \\
ADP\}
\end{array} \tag{21.35}$$

We use (21.35) to capture how a genitive with any dependency by adposition appears after the nominal head.

$$\begin{array}{l}
\%2 - Gen - bef - adp - any \\
pattern\{N[upos = NOUN]; ADP[upos = ADP]; N - > ADP; N >> \\
ADP\}
\end{array} \tag{21.36}$$

We use (21.36) to capture how a genitive with any dependency by adposition appears before the nominal head.

The first two formalizations respond to a conception of the adposition (and, by extension, the genitive it introduces) as a modifier of the noun (*mod*), which could be understood as a complement of the noun. In the first case, the noun goes before the genitive, and in the second, it goes after it. The following two formalizations respond to the conception of the relationship between the noun head and the adposition as *udep*, which is a much less free and more necessary case for the head. We consider that there are cases in which the genitive does not appear merely as a complement of the noun, but the noun itself requires this genitive. The last two formalizations respond to a dependency relation without specifying what type of dependency is between a noun and an adposition. Therefore, in this case, we would include the two previous ones (and any possibility). We must remember that to restrict any adposition to the one we are interested in by using lemma frequency, we must add the “*key*”: “*ADP.lemma*”. In all these formalizations, we have proposed the two possible orders of the genitive, before or after the noun. It is also possible to restrict by adjacency position and not in free format. However, we believe that adjacency only has a greater weight when there is no such adposition; therefore, the genitive marking is much more dependent on the position of the second noun.

About the cases of absence of adposition, we understand that the genitive is expressed by mere word order between two nouns. The one acting as dependent is the noun that would have an adposition in an equivalent language. Therefore, we will propose formalizations based on adjacent order ($<$ or $>$) in this case. However, we also open the possibility to employ it in terms of direction in general ($<<$ or $>>$), assuming that we may obtain occurrences of undesired cases.

For these cases, we understand that there are different ways of labeling this relationship between nouns that may correspond to the genitive value in a more classical nomenclature. First, the relation “N1-[*mod*]->N2” in which there is a noun that is a modifier of another noun, i.e., equivalent to what was proposed above but without the adposition as in (21.37) or (21.38). There is also the possibility of finding genitive structures in languages with two adjacent nouns using the dependencies “N1-[*compound*]->N2” or “N1-[*appos*]->N2” (compounds or appositions). However, we find that it is very complicated to draw a line to delineate these possible occurrences from other cases that could not be considered strictly genitive and can be considered, in a more orthodox way, compound words or apposition, in the original sense of the conception of these dependency labels. The patterns of these possible structures would be (21.39), (21.40), (21.41) and (21.42).

$$\begin{array}{l}
\%2 - Gen - nn - mod - n2aft - ady \\
pattern\{N1[upos = NOUN]; N2[upos = NOUN]; N1 - [mod] - > \\
N2; N1 < N2\}
\end{array} \tag{21.37}$$

We use (21.37) to capture adjacency genitive modifiers after the nominal head.

$$\begin{array}{l}
\%2 - Gen - nn - mod - n2bef - ady \\
pattern\{N1[upos = NOUN]; N2[upos = NOUN]; N1 - [mod] - > \\
N2; N1 > N2\}
\end{array} \tag{21.38}$$

We use (21.38) to capture adjacency genitive modifiers before the nominal head.

$$\begin{array}{l}
\%2 - Gen - nn - comp - n2aft - ady \\
pattern\{N1[upos = NOUN]; N2[upos = NOUN]; N1 - [compound] - > \\
N2; N1 < N2\}
\end{array} \tag{21.39}$$

We use (21.39) to capture adjacency genitive compounds after the nominal head.

$$\begin{array}{l}
\%2 - Gen - nn - comp - n2bef - ady \\
pattern\{N1[upos = NOUN]; N2[upos = NOUN]; N1 - [compound] - > \\
N2; N1 > N2\}
\end{array} \tag{21.40}$$

We use (21.38) to capture adjacency genitive compounds before the nominal head.

$$\begin{array}{l}
\%2 - Gen - nn - appos - n2aft - ady \\
pattern\{N1[upos = NOUN]; N2[upos = NOUN]; N1 - [appos] - > \\
N2; N1 < N2\}
\end{array} \tag{21.41}$$

We use (21.41) to capture adjacency genitive appositions after the nominal head.

$$\begin{array}{l}
\%2 - Gen - nn - appos - n2bef - ady \\
pattern\{N1[upos = NOUN]; N2[upos = NOUN]; N1 - [appos] - > \\
N2; N1 > N2\}
\end{array} \tag{21.42}$$

We use (21.42) to capture adjacency genitive appositions before the nominal head.

21.2.3 Trichotomous Expression Based on Grew-Match Data

The typological Universal in general (taking into account the two parts that are exposed):

1. Is fulfilled in 60% of the languages we analyzed in UD.
2. We believe that it is interesting to see if such a result arises in a balanced way in the two linguistic types that compose it. We found that the claim corresponding to postpositions has a 92% acceptance rate, while the claim referring to prepositions has a 47% acceptance rate. These data illustrate an evident imbalance.

We have explored the possibility that the trigger of this claim is the direction of the adjective, and not that of the adposition, by changing “postpositional” to “adjectives before the noun” and “prepositional” to “adjectives after the noun”. We found a worse result when substituting postpositions but a better result when changing prepositions (66%). However, we do not consider such an improvement too significant (especially since the other part of the Universal does not improve as well). This could lead to a reformulation of the Universal a: “In languages with adjectives after the noun, the genitive almost always follows the governing noun, while in languages with postpositions, it almost always precedes”. However, we believe that a process of simplification of the Universal may be more attractive, which would guarantee greater predictive power and also greater clarity: “In languages with postpositions the genitive almost always follows the governing noun”.

This Universal could be represented graphically as we propose in the figure 21.6.

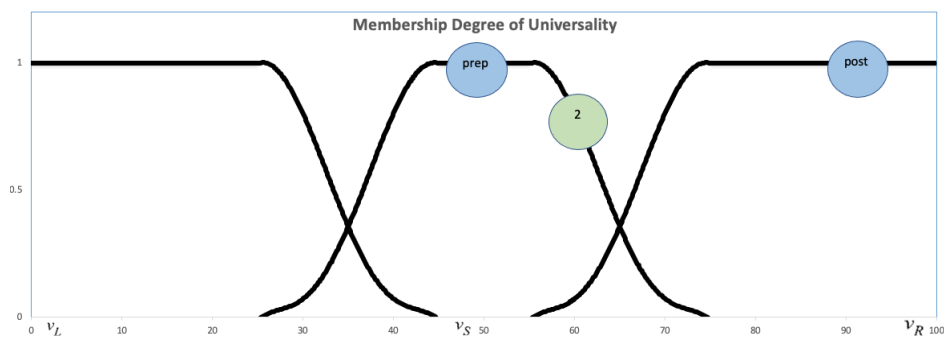


Figure 21.6: Trichotomous Expression for Universal 2

The data presented for this Universal (and 5) correspond to the cases of the linguistic type in which the genitive is marked by the case (universe of 48 languages). Although a formalization is

possible for the rest of the languages, as has been done, a confirmation of the captured cases is needed, which does not yet exist in the UD annotation by genitive marking. We believe a complete representation of the data in all languages will be possible if interglossae tagging is adopted. In Greenberg’s terms, the Universal appears as a Mid-Universal rather clearly. This is caused by the Mid-Universal representing the part relating to prepositions. The part relating to postpositions is a High-Universal and prototypical. Deleting the Mid-Universal part could turn this Universal into a High-Universal in an unquestionable way, making the claim much more reliable and valuable. We can summarise our findings as in table 21.4.

Summary of characteristics of the Universal 2 analysis:
1. Different ways of expressing the genitive. It would be useful to specify the different modes.
2. Fulfilment of 60% of the original Universal.
3. The part relating to postpositions is completed in 92%.
4. The part relating to prepositions is 47% fulfilled.
5. In its original formulation, it is a Mid-Universal.
6. Its not high level of Universality is conditioned by the Mid-Universal related to prepositions.
7. An exclusive application of the part of the postpositions is more useful and High-Universal.

Table 21.4: Summary of Characteristics of the Universal 2 Analysis

21.3 Universal 3

UNIVERSAL 3: *Languages with Dominant VSO Order are Always Prepositional.*

21.3.1 How to Check It in Typometrics

Technical Decisions:

1. Firstly, the Universal is analyzed with Typometrics (Gerdes et al., 2021).

2. We select the SUD annotation scheme because we can get results more comparable to the classical typological claims.

3. We select Graph 2D.

4. We set a filter of utterances to avoid misleading data.

5. In the x-Axis we look for *direction – cfc* of *VERB – subj – NOUN* (nouns being subjects of verbs heads).

6. In the y-Axis we look for *direction – cfc* regarding *ADP – comp : obj – NOUN* (adpositions being the head of nouns).

We have seen how there is a Universal tendency to place object after subject, so for simplifying the output, we can just take into consideration the relation between subject and verb and then check and confirm there is no mistake in using this strategy by using WALS (Dryer & Haspelmath, 2013; Haspelmath et al., 2005).

Table 21.5: Technical Decisions for Checking Universal 3 in Typometrics

We have to remember that Greenberg talks about prepositions and postpositions; it refers to the relation between an adposition and a noun since it is the most frequent and prototypical case. He was not thinking about other less frequent and more borderline cases, since it is difficult to find this information in grammars. Fortunately, we now have more fine-grained data regarding usage involving any part of speech. This means that, even if we have to check the relation with nouns because it is what Greenberg is proposing, we offer a more detailed analysis by changing the y-Axis of *ADP – comp : obj – NOUN* to other labels as could be *ADP – comp : obj – PRON* or *ADP – comp : obj – VERB*, two other common situations that presumably behave in the same

way, but that are less studied. We can also check the opposite direction: What happens when the adposition is not the head but the dependent (in SUD terms)? That is, not checking utterances like “ **Para salir** debes apretar el botón” (“To exit you must press the button” in Spanish), but checking utterances like “ Quiero **salir de** fiesta” (“I want to party” in Spanish). In this case, we have to place in the y-Axis: *VERB – comp : obj – ADP*, although we will not check this pattern. The main reason is that the results can be biased since the constraint to place the adposition after or before the verb when this one is acting as a head is heavily influenced by the fact of being an OV or VO language (Chinese, mainly Mandarin, would be a possible exception we would like to check). Then, since there are different triggers in competition (even though the VO/OV seems the most powerful), we prefer to leave these specific situations for other, more complex Universals.

However, one of the main problems in working between the classical categorical claims by Greenberg and pure quantitative data is establishing equivalences between the labels and the quantitative data. That is, we have to set a threshold according to the different words used by Greenberg. This is not an easy task since sometimes the used labels are confusing, and there can be multiple interpretations. In this study, we propose a possible threshold that seems reasonable to us based on statistical linguistic claims (confirmed by Fuzzy Logic). However, any other proposal to compute Greenberg labels could be accepted and easily applied to our data.

The first of the decisions is which languages are considered VSO and which are not (since only VSO languages will be considered in the universe of this Universal). In Typometrics, we propose a rapid way to set the threshold that can be confirmed later with classical typological data from WALS in case of doubts (Dryer & Haspelmath, 2013). We divide the graph in a trichotomous expression, that is, in 3 main sets:

1. SV
2. nondominant order
3. VS

Each of the sets takes the same proportion in the graph: SV (0-33.33), nondominant order (33.33-66.66), VS (66.66-100). This Universal applies only to languages placed in the last group (out of the 87 available languages). Then, we confirm we are correct by analyzing the languages in the VS set (Arabic, Welsh, Irish, Gaelic, and Manx). All these languages appear in WALS as VSO, so our hypothesis is confirmed. It also confirms the fact of only taking into account clear VS languages. A dichotomous classification of languages in SV-VS would have forced us to place the

threshold in the 50; this would mean we would also consider Kiche, Guajajara, Breton, and Coptic. These languages are VS in UD and SVO in WALS (no data for Kiche). Other languages, such as Tagalog, do not appear because the number of occurrences is not high enough to be 100% sure the data are representative enough. Using three sets relatively avoids these problems of consideration and categorization in the limits. Then, the universe of this Universal is 5, in our case, from the available data.

Then, we must check whether these five languages are prepositional or postpositional. For this, we follow the same methodology. Meaning that we have three main groups (in discrete terms following Gries (2013) for favoring the comparison):

1. prepositional (0-33.33)
2. no dominant order (33.33-66.66)
3. postpositional (66.66-100)

It should be noted that Greenberg uses the quantifier “always” in this Universal (and may use others in other Universals). In our study, we do not consider the nuances provided by Greenberg’s quantifiers. We believe that the work with quantifiers in Fuzzy Logic is unresolved. Moreover, it is not in our interest to assess the fulfillment of a Universal under the influence of a given quantifier. However, we should aim to describe the different phenomena without any restriction and then see in which part of our model they are located.

21.3.2 Formalization Proposed for Grew-Match

There is also the possibility to check the Universal by Grew-Match. In this case, firstly, we have to determine the basic word order of all languages to determine which have a dominant VSO order. For this, we reuse the patterns created for word order in Universal 1, section 21.1, as we show now in (21.43) - (21.48):

$ \begin{aligned} &\%3 - SVO \\ &pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = \\ &PROPN NOUN]; V - [comp : obj] - > O; O[upos = \\ &PROPN NOUN]; V >> S; S << O; V << O\} \\ &without\{V - [punct] - > P; P[lemma = “?” “!”]\} \end{aligned} $	(21.43)

In (21.43), to capture the SVO word order, we ask the subject to be before the verb and the object to be after the verb.

$$\begin{array}{l}
 \%3 - SOV \\
 pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = \\
 PROPEN|NOUN]; V - [comp : obj] - > O; O[upos = \\
 PROPEN|NOUN]; V >> S; S << O; V >> O\} \\
 without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
 \end{array} \tag{21.44}$$

In (21.44), to capture the SOV word order, we ask the subject to be before the verb and the object to be before the verb but not before the subject.

$$\begin{array}{l}
 \%3 - VSO \\
 pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = \\
 PROPEN|NOUN]; V - [comp : obj] - > O; O[upos = \\
 PROPEN|NOUN]; V << S; S << O\} \\
 without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
 \end{array} \tag{21.45}$$

In (21.45), in the VSO case, we want a verb before the subject and before the object, but the subject before the object.

$$\begin{array}{l}
 \%3 - VOS \\
 pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = \\
 PROPEN|NOUN]; V - [comp : obj] - > O; O[upos = \\
 PROPEN|NOUN]; V << S; S >> O; V << O\} \\
 without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
 \end{array} \tag{21.46}$$

In (21.46), in VOS word order, we formalize the restriction of the verb being before the subject and before the object, but the object appearing before the subject.

$$\begin{array}{l}
 \%3 - OSV \\
 pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = \\
 PROPEN|NOUN]; V - [comp : obj] - > O; O[upos = \\
 PROPEN|NOUN]; V >> S; S >> O; V >> O\} \\
 without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
 \end{array} \tag{21.47}$$

In (21.47), in the OSV order, we ask for objects appearing before subjects and verbs, but subjects before verbs.

$$\begin{array}{l}
 \%3 - OVS \\
 pattern\{V[upos = VERB]; V - [subj]- > S; S[upos = \\
 PROPEN|NOUN]; V - [comp : obj]- > O; O[upos = \\
 PROPEN|NOUN]; V << S; S >> O; V >> O\} \\
 without\{V - [punct]- > P; P[lemma = "?"|"!"]\}
 \end{array} \tag{21.48}$$

In (21.48), in OVS order, we ask for objects appearing before subjects and verbs, but verbs before subjects.

Then, we have to create two patterns to check the order of the adposition and the noun, the relation Greenberg proposed. The patterns that we propose are (21.49) and (21.50):

$$\begin{array}{l}
 \%3 - Prep - N \\
 pattern\{ADP[upos = ADP]; ADP - [comp : obj]- > N; N[upos = \\
 NOUN|PROPEN]; ADP << N\}
 \end{array} \tag{21.49}$$

In (21.49), we capture adpositions that appear before the dependent noun.

$$\begin{array}{l}
 \%3 - Post - N \\
 pattern\{ADP[upos = ADP]; ADP - [comp : obj]- > N; N[upos = \\
 NOUN|PROPEN]; ADP >> N\}
 \end{array} \tag{21.50}$$

In (21.50), we capture adpositions after the dependent noun.

In both patterns, after the label, we ask for an adposition being the head (since we are using SUD) of a noun being a dependent in a “*comp : obj*” dependency since the relation between them is not a modifying one, but the noun is an argument of the preposition in SUD terms. When looking for prepositions, we ask the adposition to precede the noun (*ADP << N*), and when looking for postpositions, we ask the adposition to be placed after the noun (*ADP >> N*). We have to remark that we formalize it in terms of word order and not in terms of word contiguity (“<<”, and not “<”) because, for example, there can be modifiers of the noun (or other examples), which means we would have another word or words between the adposition and the noun.

If we want to formalize other options worth checking, with a different part of speech or dependencies, in order to check other less common dependents of the adposition, we can propose (21.51)

- (21.56):

$$\begin{array}{l} \%3 - Prep - PN \\ pattern\{ADP[upos = ADP]; ADP - [comp : obj] - > PN; PN[upos = \\ PRON]; ADP << PN\} \end{array} \quad (21.51)$$

In (21.51), we capture adpositions that appear before the dependent pronoun.

$$\begin{array}{l} \%3 - Post - PN \\ pattern\{ADP[upos = ADP]; ADP - [comp : obj] - > PN; PN[upos = \\ PRON]; ADP >> PN\} \end{array} \quad (21.52)$$

In (21.51), we capture adpositions that appear after the dependent pronoun.

$$\begin{array}{l} \%3 - Prep - V \\ pattern\{ADP[upos = ADP]; ADP - [comp : obj] - > V; V[upos = \\ VERB]; ADP << V\} \end{array} \quad (21.53)$$

In (21.53), we capture adpositions that appear before the dependent verb.

$$\begin{array}{l} \%3 - Post - V \\ pattern\{ADP[upos = ADP]; ADP - [comp : obj] - > V; V[upos = \\ VERB]; ADP >> V\} \end{array} \quad (21.54)$$

In (21.54), we capture adpositions after the dependent verb.

$$\begin{array}{l} \%3 - Vhead - ADP \\ pattern\{ADP[upos = ADP]; V - [comp : obj] - > ADP; V[upos = \\ VERB]; ADP >> V\} \end{array} \quad (21.55)$$

In (21.55), we capture adpositions that appear after the verbal head.

$$\begin{array}{l} \%3 - ADP - Vhead \\ pattern\{ADP[upos = ADP]; V - [comp : obj] - > ADP; V[upos = \\ VERB]; ADP << V\} \end{array} \quad (21.56)$$

In (21.56), we capture adpositions that appear before the verbal head.

In the formalization (21.51) or (21.52), we are proposing the same dependency as in (21.49) or (21.50), but changing the part of speech of the dependent. In this case, we change *NOUN* for *PRON* (noun for pronoun). The rest of the elements are the same. Obviously, the difference between (21.51) and (21.52) is the word order. In (21.51), we ask for an adposition preceding the pronoun (preposition) $ADP << PN$, and in (21.52), we ask for the pronoun to precede the adposition (postposition) $ADP >> PN$. The situation in (21.53) and (21.54) is the same: we are changing the part of speech of *NOUN* to *VERB* and in (21.53), we set the order $ADP << V$ and in (21.54) we set the order $ADP >> V$. After mentioning the possibility of analyzing the situation of what happens with the dependency where an *ADP* is the head of a *VERB*, we also mentioned it would be interesting to check the behavior of *VERB* being the head of an *ADP* in the world's languages. However, as we already mentioned, checking this formalization needs to take into account other different patterns that influence this structure, and then we are not going to do it in this dissertation since it would deviate too much from the original Universal study.

The last possibility is to consider any possible part of speech being the argument of an adposition (*ADP*) in terms of SUD. To check if the correlation applies to any part of speech, the important differential factor is the dependency from this head. For this, we propose patterns (21.57) and (21.58):

%3 – <i>Prep</i> – ANY			
<i>pattern</i> { <i>ADP</i> [<i>upos</i>	=	<i>ADP</i>]; <i>ADP</i>	–
[<i>comp</i> : <i>obj</i>]–	>	ANY; ANY[<i>upos</i>	=
<i>ADJ</i> <i>ADV</i> <i>INTJ</i> <i>NOUN</i> <i>PROPN</i> <i>VERB</i> <i>ADP</i> <i>AUX</i>			
<i>CCONJ</i> <i>DET</i> <i>NUM</i> <i>PART</i> <i>PRON</i> <i>SCONJ</i> <i>SYM</i>];			
<i>ADP</i> << ANY}			

(21.57)

In (21.57), we capture adpositions that appear before any dependent part of speech.

%3 – <i>Post</i> – ANY			
<i>pattern</i> { <i>ADP</i> [<i>upos</i>	=	<i>ADP</i>]; <i>ADP</i>	–
[<i>comp</i> : <i>obj</i>]–	>	ANY; ANY[<i>upos</i>	=
<i>ADJ</i> <i>ADV</i> <i>INTJ</i> <i>NOUN</i> <i>PROPN</i> <i>VERB</i> <i>ADP</i> <i>AUX</i>			
<i>CCONJ</i> <i>DET</i> <i>NUM</i> <i>PART</i> <i>PRON</i> <i>SCONJ</i> <i>SYM</i>];			
<i>ADP</i> >> ANY}			

(21.58)

In (21.58), we capture adpositions that appear after any dependent part of speech.

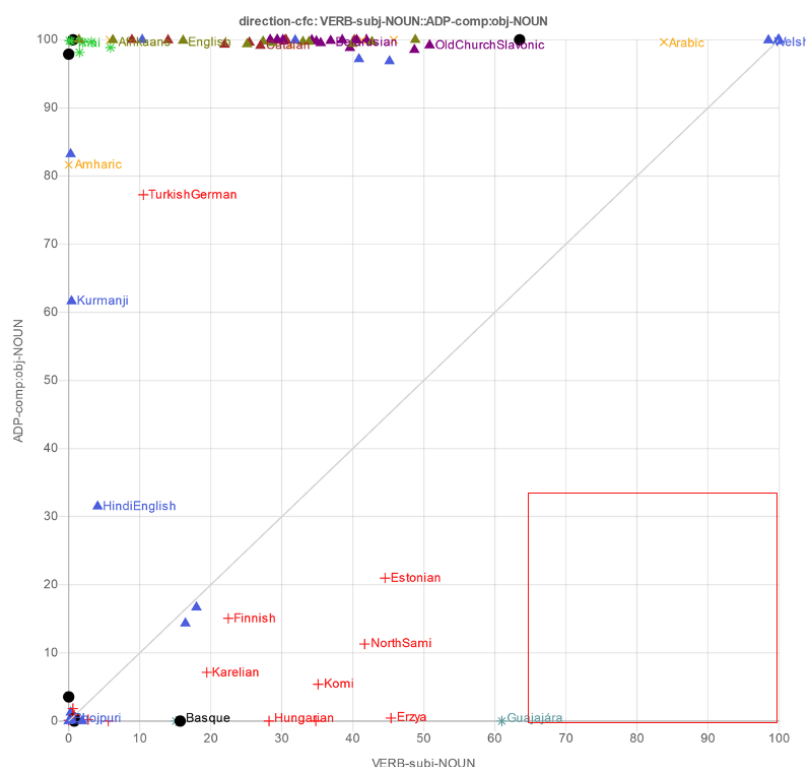


Figure 21.7: Nominal Subjects and Adpositions with Nouns Direction Checking

21.3.3 Graphs from Typometrics

As discussed in the section on this Universal, we should only consider languages of a given type on the right-hand side of the figure 21.7, corresponding to VSO. These are placed at the top, corresponding to the label “dominantly prepositional” (66.66-100), leaving the marked red square empty, as Greenberg had predicted. In this case, therefore, it seems that the linguist’s formulation is appropriate. In figure 21.8, we have chosen to highlight a C-shaped (imperfect) arrangement of languages. In this case, it can be seen how there are languages at both the extreme top (as Belarussian) and extreme bottom of the graph (almost anywhere on the x-axis), such as Basque or Hungarian, and there is also some flexibility on the left side of the y-axis. However, no occurrences along the y-axis on the right side create this kind of C. We say that it is imperfect because the lower part of the C is shorter than the upper part in relation to the restriction proposed by Greenberg, in which only prepositional languages are observed in VSO, such as Welsh or Arabic. In addition, we would like to highlight that all these languages are located at the upper end, without showing any minimal flexibility, making them strictly prepositional.

Subsequently, we can review the relationship of adpositions with pronouns. We must remember that we have chosen the nominal representation on the x-axis since, as we have seen, Greenberg’s

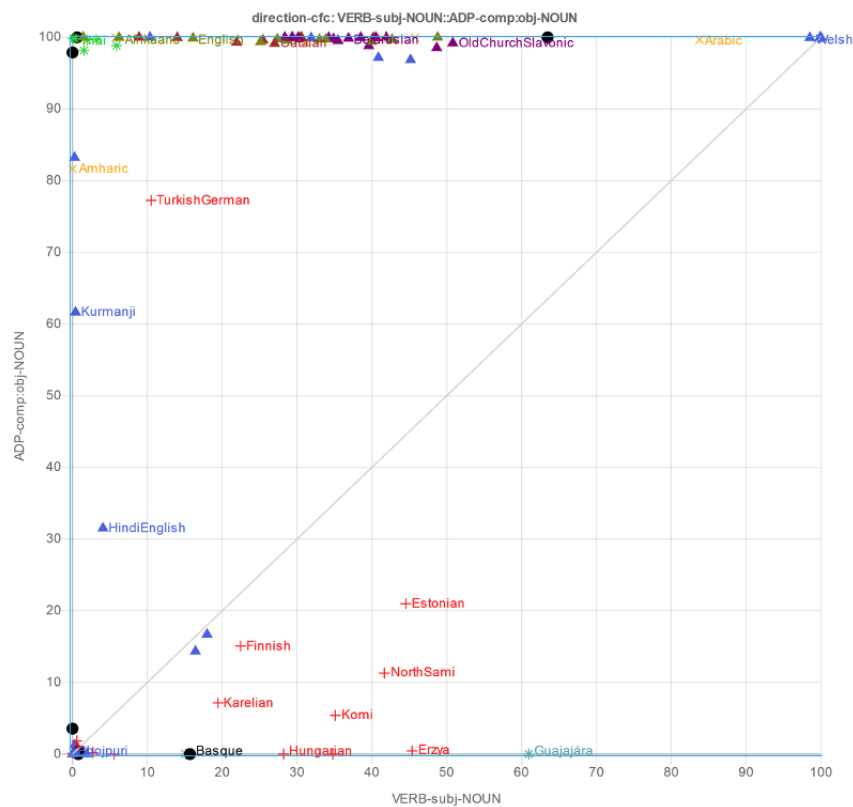
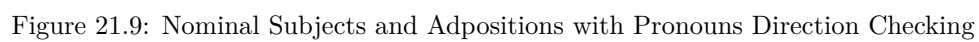


Figure 21.8: Nominal Subjects and Adpositions with Nouns Direction Proposal

proposal is the most descriptive and predictive when grouping languages in order to be consistent with the basic classical linguistic types. Thus, we observe that the VSO languages (right-hand side of figures 21.9 and 21.10) do not move and continue to place adpositions in front of the pronoun and in an equally extreme way. There is little change in this graph. As for the C-shape, we propose to represent the entire graph; this is also precisely the same as the one discussed above: there are no occurrences at the bottom or in the middle of the right-hand side of the graph.

Then, we want to look at the arrangement of adpositions with verbs as their arguments (according to how adpositions are represented in SUD), as shown in the figures 21.11 and 21.12.

In this case, we observe changes concerning the previous proposals, although in a more straightforward way in the languages that are not VSO. As for the linguistic type of the implication of Greenberg’s Universal 3, we should point out that we still find no occurrences in the red square, which allows us to extend Greenberg’s original Universal to more cases, such as this one. It should be noted that there is greater adpositional flexibility in Arabic and Irish, although these languages are still strongly prepositional in the case of verbs. In the rest of the VSO languages there is no change. In figure 21.12, it can be seen, first of all, how the lower part of the proposed “C” is even shorter. Most languages are located in the upper left triangle. This is mainly because the





languages located between 15 and 50 on the x-axis are no longer represented, as it seems they do not allow structures such as the one sought in this case.

Let us move on to verifying the Universal more comprehensively and clearly, thanks to the formalization made possible by Grew-Match. We obtain that 100% of the VSO languages in our set are prepositional. Therefore, by this means, we also confirm Greenberg's point. In the remaining cases, such as not restricting parts of speech in the syntactic elements analyzed, we still obtain a 100% success rate. We must remember that this is because we have a relatively small number of this linguistic type (VSO languages) in our set since it is not as common an order as SVO, SOV, or NDO languages, so we must wait to see what happens with such a figure once it is possible to add more languages to the study.

21.3.4 Trichotomous Expression Based on Grew-Match Data

However, if we analyze the results from the representation we are interested in, based on Fuzzy Logic, we obtain what can be seen in figure 21.13.

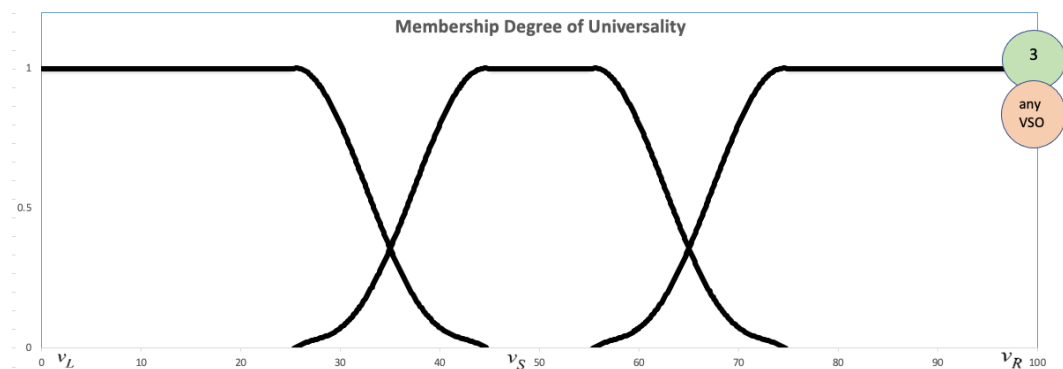


Figure 21.13: Trichotomous Expression for Universal 3

In this case, we can see that we are dealing with a clear High-Universal (and representative of its set). Therefore, we believe that, although among all the VSO languages in the world, there may be some that represent an exception to this High-Universal, it would not be compromised at all, and its validity and representativeness would be the same. On the other hand, we also want to show that the unrestricted parts of speech result of the element receiving the adposition dependency is the same as that of adpositions with nouns. Therefore, to gain representativeness of structures, we believe that a modification and extension of this Universal is interesting to remove Greenberg's implicit precision and assert that such correlation occurs not only in the canonical

cases of adposition relation but in any other. We can summarise our findings as in table 21.6.

Summary of characteristics of the Universal 3 analysis: 1. Typometrics: restriction on VS + postpositional languages (free lower right corner). 2. Typometrics: VSO languages are prepositional. 3. Typometrics: Being flexible with VS/SV order and postpositional is common in Guajajara and Finno-Ugric languages. 4. Typometrics: C-shape of Languages of the World. Restriction of variation in VS languages. 5. Typometrics: Same behavior with pronouns. 6. Typometrics: Same behavior with verbs. 7. Fulfilment of 100% of the original Universal. 8. Fulfilment of 100% of the proposal to extend to any part of speech. 9. The original Universal is a High-Universal. 10. The extension to any part of speech is a High-Universal.

Table 21.6: Summary of Characteristics of the Universal 3 Analysis

21.4 Universal 4

UNIVERSAL 4: *With overwhelmingly greater than chance frequency, languages with normal SOV order are postpositional.*

21.4.1 How to Check It in Typometrics

Technical Decisions:

1. Firstly, the Universal is analyzed with Typometrics (Gerdes et al., 2021).
2. We select the SUD annotation scheme because we can get results more comparable to the classical typological claims.
3. We select Graph 2D.
4. We set a filter of utterances to avoid misleading data.
5. In the x-Axis we select *direction – cfc* for the *VERB – comp : obj – NOUN* relation.
6. In the y-Axis we select *direction – cfc* regarding *ADP – comp : obj – NOUN*.

We have seen how there is a Universal tendency to place object after subject, so for simplifying the output, we can just take into consideration the relation between object and verb, the interesting one here, and then check and confirm there is no mistake in using this strategy by using WALS (Dryer & Haspelmath, 2013) like we have done in Universal 3, in section 21.3. We can also check the information generated by Grew-Match regarding basic word order.

Table 21.7: Technical Decisions for Checking Universal 4 in Typometrics

We have seen how there is a Universal tendency to place object after subject, so for simplifying the output, we can just take into consideration the relation between object and verb, the interesting one here, and then check and confirm there is no mistake in using this strategy by using WALS (Dryer & Haspelmath, 2013) like we have done in Universal 3, in section 21.3. We can also check the information generated by Grew-Match regarding basic word order.

We follow the same threshold as before, meaning that 0-33.33 is OV; 33.33-66.66 is nondominant order; and 66.66-100 is VO. This means that the languages we can analyze are 92. However, only 23 fit into the OV set: Afrikaans, Akkadian, Amharic, Bhojpuri, Bambara, Buryat, Basque,

Persian, Swiss German, Hindi, Western Armenian, Kazakh, Kurmanji, Korean, Marathi, Turkish German, Tamil, Telugu, Tupinamá, Turkish, Uyghur, Urdu, and Kangri. Then, we should confirm how many are SOV and not OVS or OSV in WALS. From these 23, no one is not SOV, so our hypothesis is confirmed. However, we could not consider Afrikaans, Akkadian, Swiss German, Kazakh, Kurmanji, Turkish German, Tupinamba, and Kangri because we did not find any data in WALS (grammar approach). The chances that these languages are also SOV are high, but we must be strict. Thus, we have two options. We do not consider these languages, or we can confirm the type of subject-verb-object order in these languages thanks to a parallel analysis in Grew-Match. If we ignore these languages, the universe would be 15 languages. If we perform a parallel checking in Grew-Match, we can check all 23.

In checking the basic word order of all these languages, we must confirm that all 23 candidates are part of the SOV group. The only exceptions we find are the Afrikaans and Swiss German languages, which, according to our data presented in the paper Brosa-Rodríguez and Jiménez-López (2023), in a novel approach, we indicate the truly SVO behavior of this type of Germanic languages whose categorization in basic word order is complicated due to their complex verbal distribution. This new classification proposal allows us to capture the typological behavior of these languages with an SVO structure in the rest of the syntactic relations. In addition, we would also like to point out that our proposal for checking the basic word order, much more respectful of Greenberg’s original meaning and with a higher level of detail, allows us to obtain these results. Our results are supported by detailed (minimalist) linguistic studies of Afrikaans syntax, for example (Biberauer, 2012). On the one hand, there is the traditional assignment of this type of Germanic V2 language as a variety of SOV. On the other hand, she indicates that they are, in fact, languages with a basic SVO structure, which exhibit SOV structure in specific contexts.

Furthermore, we are not able to confirm the order of Tupinamá and Kangri, as the number of occurrences of sentences with all the constraints that Greenberg requires to capture the basic word order is very low and does not allow such a check. In short, we simply eliminated these four linguistic varieties from the total 23 to be analyzed, making 19 the total number.

21.4.2 Formalization Proposed for Grew-Match

In order to check this Universal via Grew-Match, no further formalizations are necessary. Since the two elements involved are basic word order and the order of the adpositions, we can reuse the formalizations already proposed in Universal 3. In addition, we can also reuse the extra

formalizations we have already presented in U3 in connection with alternative constructions. This time, we must apply it to SOV, not VSO languages. Therefore, we reuse formalisations 21.59 - 21.66.

$$\begin{array}{l}
 \%4 - SVO \\
 pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = \\
 PROPEN|NOUN]; V - [comp : obj] - > O; O[upos = \\
 PROPEN|NOUN]; V >> S; S << O; V << O\} \\
 without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
 \end{array} \tag{21.59}$$

In (21.59), to capture the SVO word order, we ask the subject to be before the verb and the object to be after the verb.

$$\begin{array}{l}
 \%4 - SOV \\
 pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = \\
 PROPEN|NOUN]; V - [comp : obj] - > O; O[upos = \\
 PROPEN|NOUN]; V >> S; S << O; V >> O\} \\
 without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
 \end{array} \tag{21.60}$$

In (21.60), to capture the SOV word order, we ask the subject to be before the verb and the object to be before the verb but not before the subject.

$$\begin{array}{l}
 \%4 - VSO \\
 pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = \\
 PROPEN|NOUN]; V - [comp : obj] - > O; O[upos = \\
 PROPEN|NOUN]; V << S; S << O\} \\
 without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
 \end{array} \tag{21.61}$$

In (21.61), in the VSO case, we want a verb before the subject and before the object, but the subject before the object.

$$\begin{array}{l}
 \%4 - VOS \\
 pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = \\
 PROPEN|NOUN]; V - [comp : obj] - > O; O[upos = \\
 PROPEN|NOUN]; V << S; S >> O; V << O\} \\
 without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
 \end{array} \tag{21.62}$$

In (21.62), in VOS word order, we formalize the restriction of the verb being before the subject and before the object, but the object appearing before the subject.

$$\begin{array}{l}
 \%4 - OSV \\
 pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = \\
 PROPEN|NOUN]; V - [comp : obj] - > O; O[upos = \\
 PROPEN|NOUN]; V >> S; S >> O; V >> O\} \\
 without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
 \end{array} \tag{21.63}$$

In (21.63), in the OSV order, we ask for objects appearing before subjects and verbs, but subjects before verbs.

$$\begin{array}{l}
 \%4 - OVS \\
 pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = \\
 PROPEN|NOUN]; V - [comp : obj] - > O; O[upos = \\
 PROPEN|NOUN]; V << S; S >> O; V >> O\} \\
 without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
 \end{array} \tag{21.64}$$

In (21.64), in OVS order, we ask for objects appearing before subjects and verbs, but verbs before subjects.

Then, we have to create two patterns to check the order of the adposition and the noun, the relation Greenberg proposed. The patterns that we propose are (21.65) and (21.66):

$$\begin{array}{l}
 \%4 - Prep - N \\
 pattern\{ADP[upos = ADP]; ADP - [comp : obj] - > N; N[upos = \\
 NOUN|PROPEN]; ADP << N\}
 \end{array} \tag{21.65}$$

In (21.65), we want to capture adpositions that appear before the dependent noun.

$$\begin{array}{l}
 \%4 - Post - N \\
 pattern\{ADP[upos = ADP]; ADP - [comp : obj] - > N; N[upos = \\
 NOUN|PROPEN]; ADP >> N\}
 \end{array} \tag{21.66}$$

In (21.66), we want to capture adpositions after the dependent noun.

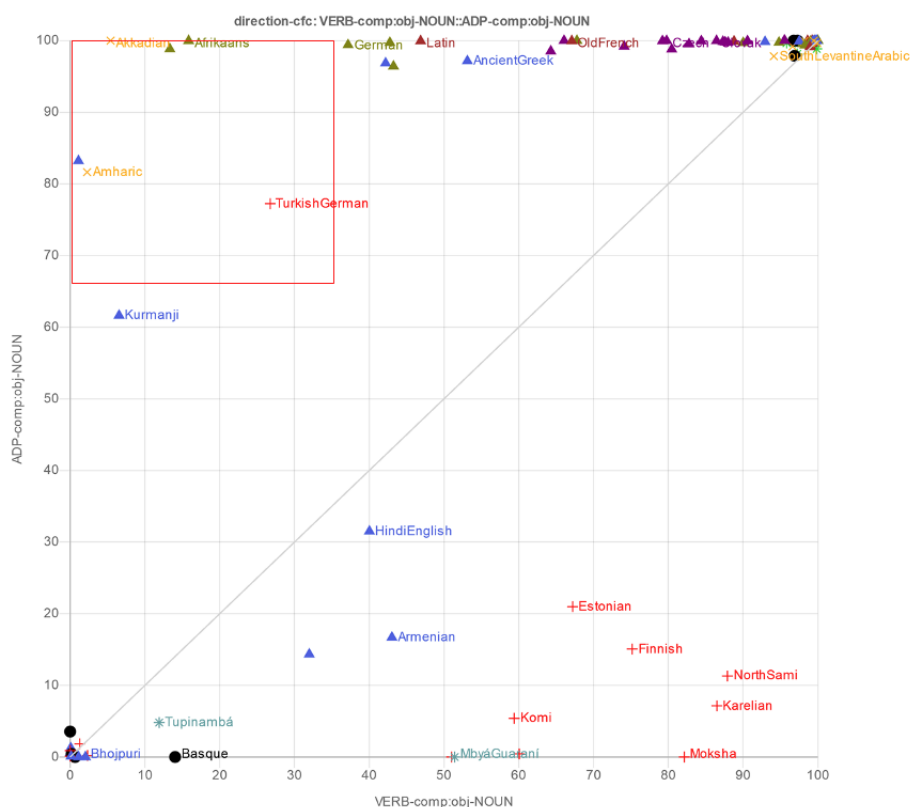


Figure 21.14: Nominal Objects and Adpositions with Nouns Direction Checking

21.4.3 Graphs from Typometrics

The fourth Universal proposed by Greenberg refers to the same aspect reviewed above but affects another linguistic type. In figures 21.14 and 21.15, we can see a Typometrics approach to the phenomenon (although, as we have always said, only the OV linguistic type is covered, and not SOV). The original and conservative result can only be refined once we present the results we will obtain in Grew-Match, thanks to our proposed formalization. However, already in the formalization, the calculations have been made to know whether the languages of both criteria coincide or whether some should be discarded/added. In this case, we must look at the languages on the left. According to Greenberg, these languages will always be located in the lower part, leaving the upper left part (where the red square is located) empty.

As can be seen, this is not the case, and we find some languages in this red square. Therefore, although there seem to be many more languages at the bottom, we do not know if the presence of these languages in such a square is relevant when validating such a Universal (until we see the data in Grew-Match). What is interesting in Typometrics is that we can see which languages appear there, whether they appear in varied/extreme positions or not, whether they are from different

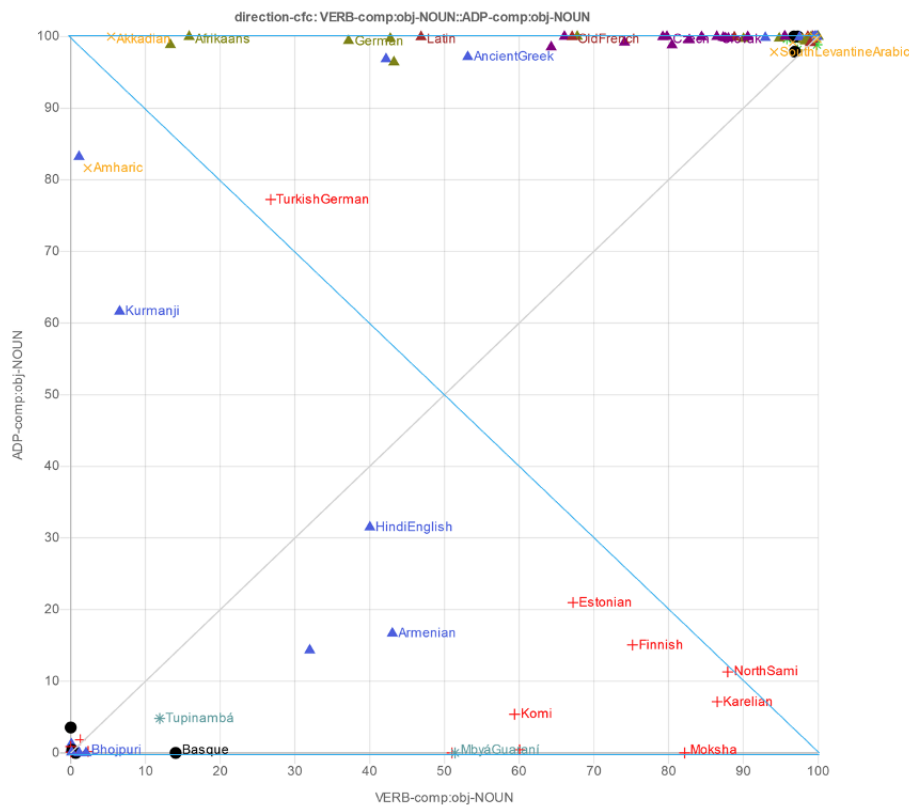


Figure 21.15: Nominal Objects and Adpositions with Nouns Direction Proposal

families or not, etc, in a quite visual way. As can be seen, the languages affected are of different origins; therefore, we believe these exceptions are more relevant than if all the languages were of the same color and shape.

If we look at the distribution of the graph in general, in figure 21.15, we will observe, first of all, that there are two very clear poles:

- the upper right part (SVO+prep)
- the lower left part (SOV+post)

Both at one pole and at the other, we observe languages moving to the opposite corners, which means that it is possible to find (although these are a clear minority):

- SVO and postpositional languages
- SOV and prepositional languages

What we notice is that what remains empty is the center of the graph, especially on the right and left sides. This implies that languages tend to be prepositional or postpositional without having

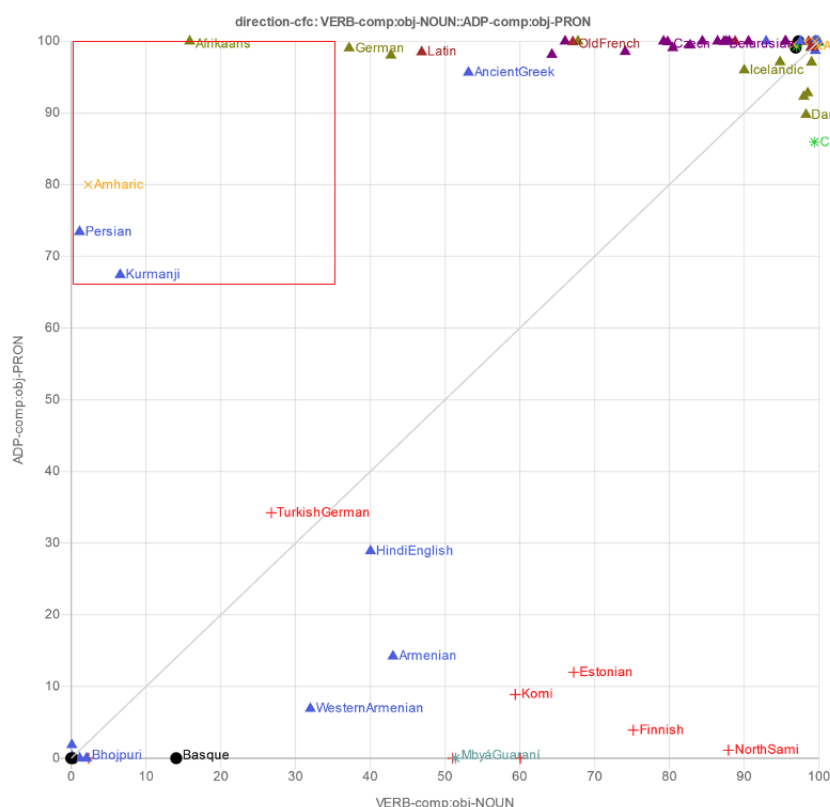


Figure 21.16: Nominal Objects and Adpositions with Pronouns Direction Checking

a non-dominant order. This gives rise to this sort of inverted Z that we believe is slightly drawn in this graph, which supports this proposal. Finally, we would also like to highlight that strict SVO languages, which are usually prepositional, can tolerate the presence of some postpositions minimally. However, some postpositional languages contain a relevant number of prepositions.

Later, we can see what happens with adpositions and pronouns in the figures 21.16 and 21.17.

In the case of the graph 21.16, in which we check whether there are languages in the place where Greenberg does not expect languages, we must say that, again, we continue to see different counterexamples (and from different areas and families). If we look, in the following graph 21.17, at the overall behavior of the languages, we will appreciate different aspects. On the one hand, we continue to see, more or less, an inverted Z-shaped distribution (so no change in general). We also see again the two main poles in the upper right and lower left corners. Instead, we want to highlight that some languages exhibit greater flexibility with the order of the adposition when it appears as a head in a dependency with a pronoun as a dependent.

If we look at the distribution by changing the dependent from noun to verb, in figure 21.18, we observe that languages continue to appear in the red square where they should not appear.

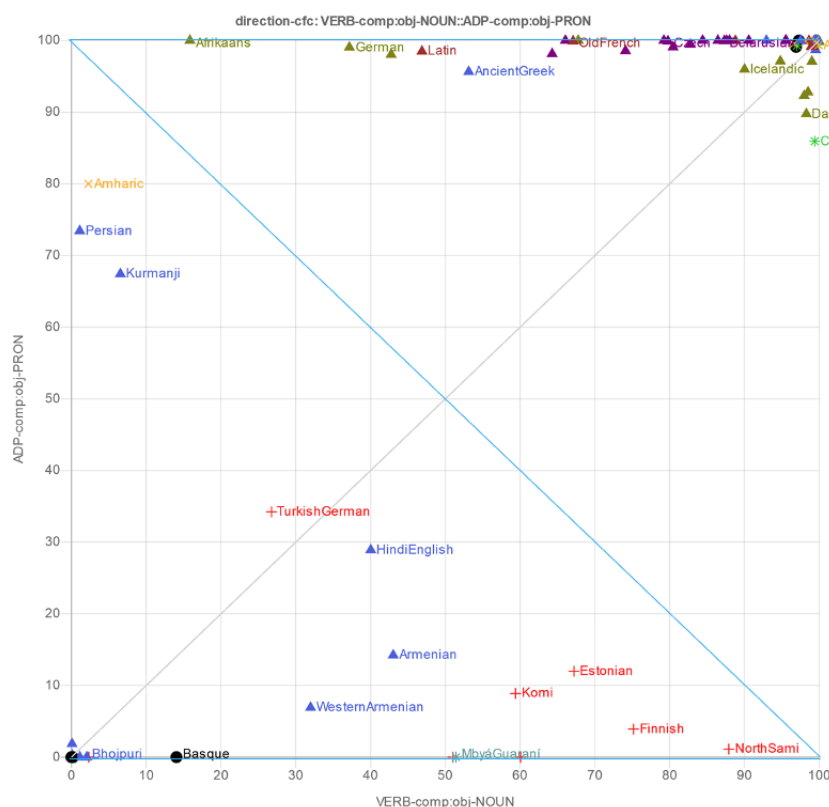


Figure 21.17: Nominal Objects and Adpositions with Pronouns Direction Proposal

However, no precise distribution of patterns in the general graph is apparent. We believe this is due mainly to a lower presence of languages with this type of structure, which causes the graph to change its representation.

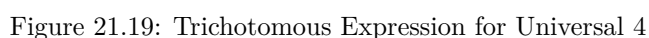
21.4.4 Trichotomous Expression Based on Grew-Match Data

The main results we have detected are the following:

- Regarding the exact quantitative result of compliance with Greenberg's Universal 4, we obtain a 71% compliance rate.
- In this case, we also want to show the result of a Universal proposed by us based on Greenberg's formulation, in which we check the relationship between the dominant basic order SVO and the presence of prepositions (i.e., coincidence with the linguistic type VSO). In this case, we obtain an 83% success rate, clearly higher than Greenberg's proposal.
- It is also possible to check what happens in the original Universal and in our proposal when we remove the canonical structure restriction of adpositions in Typology (adposition + noun)



This, taken to the model proposed in this thesis, would be as in figure 21.19.



As can be seen, the Universal proposed by Greenberg is a High-Universal, but it does not appear in the high value of membership within the group itself. This means that it is a Universal that

must be considered since its predictive value is relevant and it is part of such a group. However, it would not be a canonical or typical example of this group since it does not exhibit the maximum value of this group. On the other hand, the new Universal we propose, relating SVO languages to prepositions, would be highly High-Universal and representative of its group. Likewise, the proposals of both linguistic types (SOV and SVO) without any parts of speech restriction on the nominal dependents of both parts of implication are also High-Universal and representative. Therefore, especially in order to extend the Universal as much as possible and gain predictive power by not restricting it exclusively to the typical and representative part of speech (adposition + noun), we believe that these new proposals offer more interesting information without contradicting or ignoring the original nominal dependency, which is still of great value. We can summarise our findings as in table 21.8.

Summary of characteristics of the Universal 4 analysis:
1. Typometrics: exception for some Germanic or Semitic languages (upper left corner with languages).
2. Typometrics: Inverted Z pattern. Restriction on most central positions.
3. Typometrics: same behavior with pronouns, exceptions, and pattern form.
4. Typometrics: Still exceptions with verbs. There is no longer a clear pattern of distribution form.
5. Fulfilment of the original Universal in 71%.
6. 83% compliance with the SVO version of the Universal.
7. Compliance of SVO or SOV languages with any part of speech by 81%.
8. The original Universal is a non-prototypical High-Universal. Future checks should be made to see how it evolves.
9. The proposed alternative Universal is a prototypical High-Universal. It is more reliable.

Table 21.8: Summary of Characteristics of the Universal 4 Analysis

21.5 Universal 5

UNIVERSAL 5: *If a language has dominant SOV order and the genitive follows the governing noun, the adjective likewise follows the noun.*

21.5.1 How to Partially Check it in Grew-Match

This Universal contains two implications that must be fulfilled to find the result presented in the conclusion. The first of these necessary implications is having the dominant order SOV. For this, as we have already pointed out in Universal 1, for example, it is sufficient to apply the patterns (21.1) - (21.6).

%5 – SVO

$pattern\{V[upos = VERB]; V - [subj] - > S; S[upos =$
 $PROPN|NOUN]; V - [comp : obj] - > O; O[upos =$
 $PROPN|NOUN]; V >> S; S << O; V << O\}$
 $without\{V - [punct] - > P; P[lemma = "?"|"!"]\}$

(21.67)

We use (21.67) to capture the SVO word order. We ask the subject to be before the verb and the object to be after the verb.

%5 – SOV

$pattern\{V[upos = VERB]; V - [subj] - > S; S[upos =$
 $PROPN|NOUN]; V - [comp : obj] - > O; O[upos =$
 $PROPN|NOUN]; V >> S; S << O; V >> O\}$
 $without\{V - [punct] - > P; P[lemma = "?"|"!"]\}$

(21.68)

We use (21.68) to capture the SOV word order. We ask the subject to be before the verb and the object before the verb but not before the subject.

%5 – VSO

$pattern\{V[upos = VERB]; V - [subj] - > S; S[upos =$
 $PROPN|NOUN]; V - [comp : obj] - > O; O[upos =$
 $PROPN|NOUN]; V << S; S << O\}$
 $without\{V - [punct] - > P; P[lemma = "?"|"!"]\}$

(21.69)

In (21.69), in the VSO case, we want a verb before the subject and before the object, but the subject before the object.

$$\begin{array}{l}
 \%5 - VOS \\
 pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = \\
 PROPEN|NOUN]; V - [comp : obj] - > O; O[upos = \\
 PROPEN|NOUN]; V << S; S >> O; V << O\} \\
 without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
 \end{array} \quad (21.70)$$

In (21.70), in VOS word order, we formalize the restriction of the verb being before the subject and before the object, but the object appearing before the subject.

$$\begin{array}{l}
 \%5 - OSV \\
 pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = \\
 PROPEN|NOUN]; V - [comp : obj] - > O; O[upos = \\
 PROPEN|NOUN]; V >> S; S >> O; V >> O\} \\
 without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
 \end{array} \quad (21.71)$$

In (21.71), in the OSV order, we ask for objects appearing before subjects and verbs, but subjects before verbs.

$$\begin{array}{l}
 \%5 - OVS \\
 pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = \\
 PROPEN|NOUN]; V - [comp : obj] - > O; O[upos = \\
 PROPEN|NOUN]; V << S; S >> O; V >> O\} \\
 without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
 \end{array} \quad (21.72)$$

In (21.72), in OVS order, we ask for objects appearing before subjects and verbs, but verbs before subjects.

Once we obtain the occurrences, filter and process them as we have already shown in several previous Universals, we calculate the dominant word order by establishing a ranking order by number of occurrences and assigning the value “dominant SOV order” to the languages whose highest value is SOV and is double or higher than the second order found. Once we have that list with SOV languages, we must determine how many have the genitive after the noun it depends on. To do this, we refer to the approach presented in Universal 2. Again, given the difficulties of the subject and the limitations encountered in UD, as an example, we develop how this Universal would

be checked in languages that express the genitive by case, taking advantage of the formalizations (21.29) and (21.30). Therefore, we note that, again, the testing of this Universal is more limited than desired, and we hope to be able to expand the languages analyzed in the future once we test the more reliable way of taking into account other ways of expressing the genitive.

$$\begin{array}{l} \%5 - Gen - aft - adp \\ pattern\{N[upos = NOUN]; ADP[upos = ADP]; N - [mod] - > \\ ADP; N << ADP\} \end{array} \quad (21.73)$$

In (21.73), we capture how a modifying genitive by adposition appears after the nominal head.

$$\begin{array}{l} \%5 - Gen - bef - adp \\ pattern\{N[upos = NOUN]; ADP[upos = ADP]; N - [mod] - > \\ ADP; N >> ADP\} \end{array} \quad (21.74)$$

In (21.74), we capture how a modifying genitive by adposition appears before the nominal head.

Once we have confirmed the languages that have SOV dominant order and that place the genitive after the noun head (at least through the case, although it is undoubtedly the most recurrent mode in many SOV languages), we must check the order of the adjective as modifier of the noun, which must be, like the genitive, in second position.

$$\begin{array}{l} \%5 - ADJ - after \\ pattern\{ADJ[upos = ADJ]; N - [mod] - > ADJ; N[upos = \\ PROPEN|NOUN]; N << ADJ\} \end{array} \quad (21.75)$$

(21.75) corresponds to adjectives after modified nouns.

$$\begin{array}{l} \%5 - ADJ - before \\ pattern\{ADJ[upos = ADJ]; N - [mod] - > ADJ; N[upos = \\ PROPEN|NOUN]; N >> ADJ\} \end{array} \quad (21.76)$$

(21.76) corresponds to adjectives before modified nouns.

In short, we only review languages whose dominant order is SOV and place their genitive after the noun head (at the moment, only those with cases for it), and if they place the adjective after the modified noun, the ascription to the Universal will be correct. In contrast, if the adjective is placed before the noun head, it will be an exception to the Universal.

21.5.2 Trichotomous Expression Based on Grew-Match Data

Following the guidelines of the Universal, we obtain a compliance of 100% of the Universal, which is a misleading percentage. The representativeness is not adequate since only one language meets these requirements. We have only obtained one language with a dominant nominal SOV order, which has genitive cases after the noun as the dominant order, a precondition to search for the direction of the adjective. In most cases, we deal with languages that place the genitive in another position or do not exhibit a dominant order in any of the elements. Therefore, we only offer this data as an anecdotal point of reference that may be relevant for future studies in which more languages are added and contrast is allowed.

The most plausible possibility is that Greenberg has made a combination of two relatively rare types that cause this anomaly in the available languages. Therefore, we wanted to further check other possible combinations, which could point to correlations in other senses by playing with the same elements and being more representative of the usual type combinations of the languages. We wanted to check the SOV type languages where the genitive case precedes the noun. We hypothesize that we would find more languages with such a combination, which seems more usual, and that successful implication would take place with a direction of modifier adjectives preceding the noun. In this case, we have also achieved 88% compliance with the Universal, with a universe of 8 languages (only one exception). Although the universe is still low, these results allow us to speak of significant results, especially when we want to check whether, with a different sample, the probabilities that the set would be different in the trichotomous expression. There is a clear trend, which can be confirmed with more languages.

In addition, we wanted to review the behavior of another linguistic type: languages with SVO dominant nominal order and with the genitive case preceding the noun. In this implication, the universe is 12 languages. Although it may also seem like a small universe, in this case, the results are statistically significant, and it is a relatively rich sample if we consider the multiple restrictions in the implication prior to testing. In this case, we obtain that 73% of the languages of this type place the adjective before the noun.

If the basic word order constraint is removed from the Universal and we change the direction of the Universal and create a statement like “If in a language the genitive precedes the governing noun, then the adjective likewise precedes the noun”, we get 79% compliance. In this case, we have a wider universe (19 languages) and a much simpler and less restrictive formulation. The results are still just as interesting from the Fuzzy Logic point of view, as shown in figure 21.20.

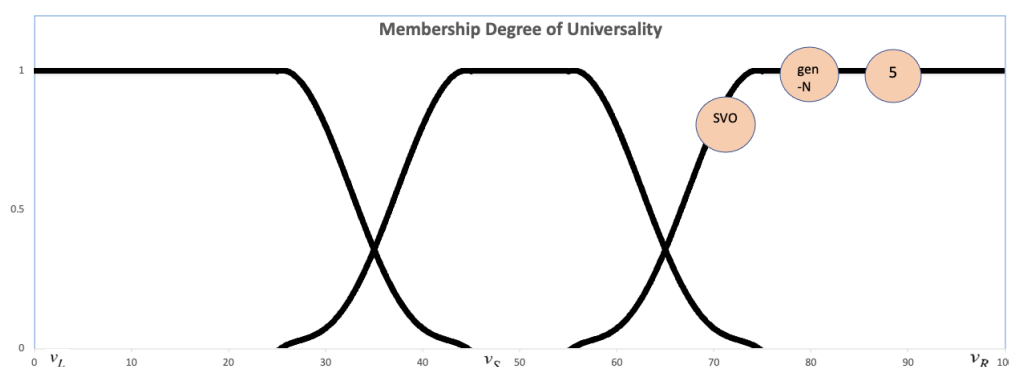


Figure 21.20: Trichotomous Expression for Universal 5

The new proposal we make of Universal 5, changing the direction of the genitive case of the SOV languages being analyzed, gives rise to the new Universal: “If a language has dominant SOV order and the genitive precedes the governing noun, then the adjective likewise precedes the noun”. This appears as a High-Universal representative (1) of its group. On the other hand, the proposition that refers to the possible Universal “If a language has dominant SVO order and the genitive precedes the governing noun, the adjective precedes the noun” is a High-Universal. In this case, its prototypicality value is not the highest possible, although it is pretty high. In the case of the last Universal we have created, taking only consistency in the direction of genitive cases and preceding nominal modifier adjectives, we also find a High-Universal representative of its set.

Therefore, the three proposals we have carried out seem more interesting than those of Greenberg. In this case, it is not so much due to a low level of compliance with the original proposal but to the impossibility of being confident of its representativeness because it only shows one language from our sample. Given that the formulated combination does not seem to be the most common in languages of this type, by expanding the sample, we do not expect many more occurrences.

In the case of our proposals, both the version that looks for the most coherent order of the SOV type, the version that characterizes the SVO order, and the version that eliminates the basic word order and focuses only on the genitive cases, we find results within the High-Universal group that improve the treatment of the adjective order prediction, even though it is complex in Typology.

We can summarise our findings as in table 21.9.

Summary of characteristics of the universal 5 analysis: 1. Compliance with the original universal in case-marked genitives is 88%. 2. The compliance with the distribution of case-marked genitive preceding the noun in SOV languages and the adjective also preceding is 79%. 3. The compliance with the distribution of case-marked genitives preceding the noun in SVO languages and the adjective also preceding is 73%. 4. The versions announced in the three previous points are High-universals. 5. The universal proposed by Greenberg is the most reliable of all the versions with the data we have been able to use.

Table 21.9: Summary of Characteristics of the Universal 5 Analysis

21.6 Universal 6

UNIVERSAL 6: *All languages with dominant VSO order have SVO as an alternative or as the only alternative basic order.*

21.6.1 Formalization Proposed for Grew-Match

This Universal states that languages with VSO order as the most frequent one will also exhibit utterances of SVO order in other situations. For this, we need to check all available languages first and decide which of them are VSO as a dominant order. For this purpose, we will use the previous criteria based on Gries (2013). Then, we need to check the remaining occurrences and determine whether there are also situations in which the basic SVO order is used exclusively or among others.

The formalizations to be used are those already described in Universal 1, (21.1). Similarly, we can reuse all the formalizations made to capture variations of the Universal (taking pronouns, any part of speech, etc.) to see if their formulation is extensible. Therefore, we do not believe repeating the formalizations and commenting on them is necessary.

As mentioned, if we obtain less than 10 occurrences, we consider that representativeness is not assured. We discard the analysis of such a language due to an insufficient number of tokens as if we did not have the information of such a language. Once this filter has been passed, any occurrence,

no matter how small (1%, for example), is already considered as a valid alternative order of a language.

$$\begin{array}{l}
 \%6 - SVO \\
 pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = \\
 PROPEN|NOUN]; V - [comp : obj] - > O; O[upos = \\
 PROPEN|NOUN]; V >> S; S << O; V << O\} \\
 without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
 \end{array} \tag{21.77}$$

To capture the SVO word order, in (21.77), we ask the subject to be before the verb and the object to be after the verb.

$$\begin{array}{l}
 \%6 - SOV \\
 pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = \\
 PROPEN|NOUN]; V - [comp : obj] - > O; O[upos = \\
 PROPEN|NOUN]; V >> S; S << O; V >> O\} \\
 without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
 \end{array} \tag{21.78}$$

To capture the SOV word order, in (21.78), we ask the subject to be before the verb and the object to be before the verb but not before the subject.

$$\begin{array}{l}
 \%6 - VSO \\
 pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = \\
 PROPEN|NOUN]; V - [comp : obj] - > O; O[upos = \\
 PROPEN|NOUN]; V << S; S << O\} \\
 without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
 \end{array} \tag{21.79}$$

In the VSO case, in (21.79), we want a verb before the subject and before the object, but the subject before the object.

$$\begin{array}{l}
 \%6 - VOS \\
 pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = \\
 PROPEN|NOUN]; V - [comp : obj] - > O; O[upos = \\
 PROPEN|NOUN]; V << S; S >> O; V << O\} \\
 without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
 \end{array} \tag{21.80}$$

In VOS word order, in (21.80), we formalize the restriction of verb being before the subject and

before the object, but the object appearing before the subject.

$$\begin{array}{l}
 \%6 - OSV \\
 pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = \\
 PROPEN|NOUN]; V - [comp : obj] - > O; O[upos = \\
 PROPEN|NOUN]; V >> S; S >> O; V >> O\} \\
 without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
 \end{array} \tag{21.81}$$

In the OSV order, in (21.81), we ask for objects appearing before subjects and verbs, but subjects before verbs.

$$\begin{array}{l}
 \%6 - OVS \\
 pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = \\
 PROPEN|NOUN]; V - [comp : obj] - > O; O[upos = \\
 PROPEN|NOUN]; V << S; S >> O; V >> O\} \\
 without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
 \end{array} \tag{21.82}$$

Moreover, in OVS order, in (21.82), we ask for objects appearing before subjects and verbs, but verbs before subjects.

21.6.2 Trichotomous Expression Based on Grew-Match Data

The quantitative results we obtained once all the verifications had been carried out show 100% compliance with the Universal. However, we must point out that in our sample, we have a representation of 4 languages. Checking the significance of this 4-0 result gives us a figure between 4 and 5 to be able to ensure the results. Moreover, we are not in a case in which genetic and area independence of the four available languages is guaranteed, which makes such reliability difficult.

We also wanted to carry out tests with other types of configurations. We especially emphasize that we do not restrict by part of speech. We must remember that Greenberg, when determining his basic word order, always relies on nominal structures. If we remove such a restriction and simply check the syntactic dependencies of basic word order without taking into account which part of speech is one of the dependents, we also obtain 100% compliance with the Universal.

Thus, we cannot clearly state that the results we show would be the same or similar to any other sample. However, we believe that in a model such as the one we offer, the results would still be in the same set as shown in figure 21.21.

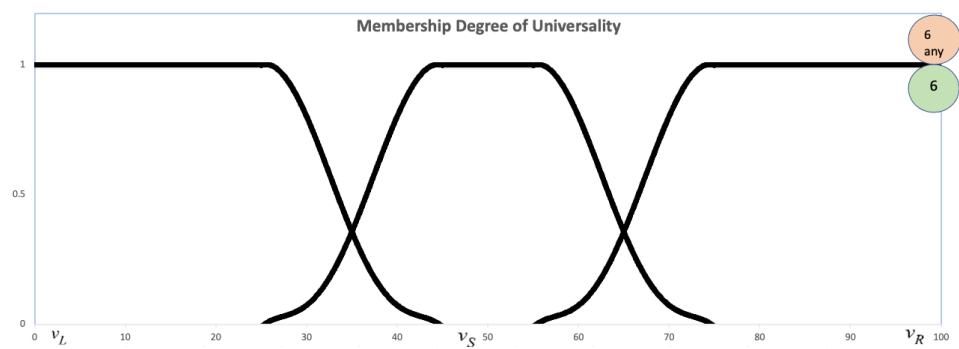


Figure 21.21: Trichotomous Expression for Universal 6

In this case, we can clearly see that in both cases, we encounter a prototypical High-Universal. Therefore, in this case, both ways can be explored; the nominal restriction is not necessary. We can summarise our findings as in table 21.10.

Summary of characteristics of the Universal 6 analysis:

1. The number of VSO languages in the world is low and causes both Greenberg’s and our sample not to present a large number of languages for analysis. Further testing should be done when more languages are available.
2. 100% of the analyzed languages meet Greenberg’s original Universal.
3. 100% of the languages analyzed to meet the Universal if there were no part of speech restriction.
4. Greenberg’s original Universal is a High-Universal.
5. The Universal with the elimination of the part of speech restriction is a High-Universal.

Table 21.10: Summary of Characteristics of the Universal 6 Analysis

21.7 Universal 7

UNIVERSAL 7: *If in a language with dominant SOV order, there is no alternative basic order or only OSV as the alternative, then all adverbial modifiers of the verb likewise precede the verb.*

21.7.1 How to partially check it in Typometrics

We have decided to place this Universal in this section because the second part is easily testable via Typometrics, which is quite interesting. However, the first part does not allow it, so we must combine these different methods.

First, we are asked, again, to analyze SOV-order languages. In this case, we do not recommend analyzing the OV or VO direction in Typometrics as we have proposed in a previous Universal because this would not be enough, as we also ask for specific behavior with the alternative orders. Therefore, we recommend entirely working on this first part of the Universal, the implication, through Grew-Match. We must, again, as in the previous Universal (21.5), apply the patterns related to the basic word order: (21.1) - (21.6).

Subsequently, we determine the dominant order using the same methodology as always, and we are left with the SOV languages only. However, we must determine an extra aspect about the other orders in this case. We only consider languages whose only order is SOV (and no alternative order exists) or with only the alternative order OSV. Otherwise, even if they are SOV languages, we must discard them for consideration in this Universal.

The list of SOV languages is Akkadian, Akuntsu, Amharic, Basque, Buryat, Hindi, Kazakh, Korean, Kurmanji, Marathi, Persian, Sanskrit, Tamil, Telugu, Turkish, Turkish German, Uyghur, Xibe, Yakut. However, the only ones that meet the extra requirement proposed by Greenberg are Amharic, Buryat, Kazakh, Korean, Marathi, Tamil, Telugu, Turkish, and Xibe.

In this case, we have focused exclusively on showing the verification pathway. Since the results are incomplete, we do not evaluate graphs in Typometrics. However, if we would like to check it, we propose the following steps.

Technical Decisions:

1. Firstly, the Universal is analyzed with Typometrics (Gerdes et al., 2021).
2. We select the SUD annotation scheme because we can get results more comparable to the classical typological claims.
3. We select Graph 2D.
4. We set a filter of utterances to avoid misleading data.
5. We can select graph 1D, then *head – initiality – cfc* and in the x-Axis locate *VERB – mod – ADV*, which corresponds to adverbs modifying verbs.
6. On the other hand, we have the option to select a 2D graph with two values *head – initiality – cfc*; in the x-Axis, we place again *VERB – mod – ADV*, and in the y-Axis, we place *VERB – comp : obj – NOUN*.

This will allow us to have a cluster with most SOV languages in a particular area (in this case, we expect to find them in the lower left corner and not in the lower right corner). We will simply discard the languages that are either not technically SOV with basic word order or do not fit the alternative order constraint.

Table 21.11: Technical Decisions for Checking Universal 7 in Typometrics

21.7.2 Formalization Proposed for Grew-Match

We obtain again the basic word order through 21.83 - 21.88.

$$\begin{aligned}
 &\%7 - SVO \\
 &pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = \\
 &PROPN|NOUN]; V - [comp : obj] - > O; O[upos = \\
 &PROPN|NOUN]; V >> S; S << O; V << O\} \\
 &without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
 \end{aligned} \tag{21.83}$$

To capture the SVO word order, in (21.83), we ask the subject to be before the verb and the object to be after the verb.

$$\begin{aligned}
 &\%7 - SOV \\
 &pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = \\
 &PROPN|NOUN]; V - [comp : obj] - > O; O[upos = \\
 &PROPN|NOUN]; V >> S; S << O; V >> O\} \\
 &without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
 \end{aligned} \tag{21.84}$$

To capture the SOV word order, in (21.84), we ask the subject to be before the verb and the object to be before the verb but not before the subject.

$$\begin{array}{l}
 \%7 - VSO \\
 pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = \\
 PROPEN|NOUN]; V - [comp : obj] - > O; O[upos = \\
 PROPEN|NOUN]; V << S; S << O\} \\
 without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
 \end{array} \tag{21.85}$$

In the VSO case, in (21.85), we want a verb before the subject and before the object, but the subject before the object.

$$\begin{array}{l}
 \%7 - VOS \\
 pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = \\
 PROPEN|NOUN]; V - [comp : obj] - > O; O[upos = \\
 PROPEN|NOUN]; V << S; S >> O; V << O\} \\
 without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
 \end{array} \tag{21.86}$$

In VOS word order, in (21.86), we formalize the restriction of verb being before the subject and before the object, but the object appearing before the subject.

$$\begin{array}{l}
 \%7 - OSV \\
 pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = \\
 PROPEN|NOUN]; V - [comp : obj] - > O; O[upos = \\
 PROPEN|NOUN]; V >> S; S >> O; V >> O\} \\
 without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
 \end{array} \tag{21.87}$$

In the OSV order, in (21.87), we ask for objects appearing before subjects and verbs, but subjects before verbs.

$$\begin{array}{l}
 \%7 - OVS \\
 pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = \\
 PROPEN|NOUN]; V - [comp : obj] - > O; O[upos = \\
 PROPEN|NOUN]; V << S; S >> O; V >> O\} \\
 without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
 \end{array} \tag{21.88}$$

Furthermore, in OVS order, in (21.88), we ask for objects appearing before subjects and verbs, but verbs before subjects.

In order to check the part done in Typometrics through Grew-Match, we simply need to propose patterns (21.89) and (21.90).

$$\begin{array}{l} \%7 - ADV - beforeV \\ pattern\{V[upos = VERB]; V - [mod]- > ADV; ADV[upos = \\ ADV]; V >> ADV\} \end{array} \quad (21.89)$$

In (21.89), we capture adverbs that precede the verbs they modify.

$$\begin{array}{l} \%7 - ADV - afterV \\ pattern\{V[upos = VERB]; V - [mod]- > ADV; ADV[upos = \\ ADV]; V << ADV\} \end{array} \quad (21.90)$$

In (21.90), we capture adverbs that appear after the verbs they modify.

In these formalizations, we create two elements, one verbal and the other adverbial. We create a dependency between them in which the element *V* (verb) is the head of a modification made by the dependent *ADV* (adverb). Each pattern corresponds to each of the two possible orders: the adverb precedes ($V >> ADV$) or the verb precedes ($V << ADV$). In addition, as usual, we can review the behavior of much less canonical but equally interesting dependencies about the adverb as a modifier. For example, we can check the behavior of adverbs modifying adjectives, adverbs modifying other adverbs, adverbs modifying any part of speech, and adverbial clauses modifying verbs. We propose formalizations such as (21.91) - (21.96) for these cases.

$$\begin{array}{l} \%7 - ADV - beforeADJ \\ pattern\{ADJ[upos = ADJ]; ADJ - [mod]- > ADV; ADV[upos = \\ ADV]; ADJ >> ADV\} \end{array} \quad (21.91)$$

In (21.91), we capture adverbs that precede the adjectives they modify.

$$\begin{array}{l} \%7 - ADV - afterADJ \\ pattern\{ADJ[upos = ADJ]; ADJ - [mod]- > ADV; ADV[upos = \\ ADV]; ADJ << ADV\} \end{array} \quad (21.92)$$

In (21.92), we capture adverbs that appear after the adjectives they modify.

$$\begin{array}{l}
\%7 - ADV - before ANY \\
pattern\{ANY[upos = ADJ|ADV|INTJ|NOUN|PROPN| \\
VERB|ADP|AUX|CCONJ|DET|NUM|PART|PRON| \\
SCONJ|SYM]; ANY - [mod]- > ADV; ADV[upos = \\
ADV]; ANY >> ADV\}
\end{array} \tag{21.93}$$

In (21.93), we capture adverbs that precede any part of speech they modify.

$$\begin{array}{l}
\%7 - ADV - after ANY \\
pattern\{ANY[upos = ADJ|ADV|INTJ|NOUN|PROPN| \\
VERB|ADP|AUX|CCONJ|DET|NUM|PART|PRON| \\
SCONJ|SYM]; ANY - [mod]- > ADV; ADV[upos = \\
ADV]; ANY << ADV\}
\end{array} \tag{21.94}$$

In (21.94), we capture adverbs that appear after any part of speech they modify.

$$\begin{array}{l}
\%7 - ADVcl - before \\
pattern\{SCONJ[upos = SCONJ]; VIF[upos = \\
VERB|AUX]; VMAIN[upos = VERB|AUX]; VMAIN - [mod]- > \\
SCONJ; SCONJ - [comp : obj]- > VIF; SCONJ << VMAIN\}
\end{array} \tag{21.95}$$

In (21.95), we capture adverbial clauses that appear before the verb they modify.

$$\begin{array}{l}
\%7 - ADVcl - after \\
pattern\{SCONJ[upos = SCONJ]; VIF[upos = \\
VERB|AUX]; VMAIN[upos = VERB|AUX]; VMAIN - [mod]- > \\
SCONJ; SCONJ - [comp : obj]- > VIF; SCONJ >> VMAIN\}
\end{array} \tag{21.96}$$

In (21.96), we capture adverbial clauses that appear after the verb they modify.

In the case of adverbs modifying adjectives, we must simply change the verbal element *V* for an adjectival element *ADJ* of the previous formalization. To analyze any part of speech, we name an element called *ANY* that corresponds to all possible parts of speech, as we have already done in previous formalizations. In order to capture adverbial clause structures, we must propose a more complex structure, given the specificity of the query. We activate three elements: a subordinator element *SCONJ*, a verb for the adverbial clause *VIF*, and a verb in the main clause *VMAIN*. On the one hand, we ask for the appearance of two dependencies: a relation between the main

verb and the subordinator element as modification *mod* and a relation between the subordinator element and the verb of this adverbial sentence as argument *comp : obj*.

It should be noted that, given the specificity of the structure, it is possible that some languages do not show occurrences since some languages may have some type of structure with a particle such as a head or in which the annotators have assigned another part of speech to the central element of the adverbial sentence as it does not correspond 100% with the structure of most European languages. However, this approach is the one we have found to be the most transversal and Universal.

21.7.3 Trichotomous Expression Based on Grew-Match Data

After restricting the possible types stated in the implication of this Universal, we obtain 89% compliance. We also wanted to check other cases, such as the case of a Universal, such: “If in a language with dominant SVO order, there is no alternative basic order, then all adverbial modifiers of the verb likewise precede the verb.” In this formulation, the fulfillment would be 64%.

On the other hand, we can also check what happens if we remove the restriction that there is no alternative order or allow only OSV. In such a case, SOV type languages show 85% following the Universal and SVO type languages show 53%.

Finally, we also wanted to modify the original Universal in the second part of the formulation concerning adverbs as verb modifiers and their direction. In the case of revising SOV languages (without alternative basic order or only OSV), in 85% of the cases, they place the adverb as a modifier of any part of speech before the element they modify. In the case of adverbs modifying adjectives, they are placed in front of their head in 100% of the cases.

All these different formulations are shown graphically in figure 21.22:

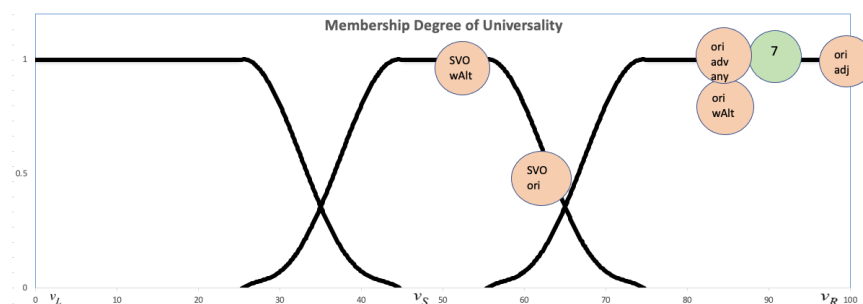


Figure 21.22: Trichotomous Expression for Universal 7

The original Universal formulated by Greenberg is placed in the High-Universal group with a maximum prototypicality of such a group. Proposals in which the requirements of not having alternative orders are eliminated or the relation of adverbs with adjectives or any part of speech (and not with verbs) are checked also have the same consideration. To gain greater representativeness and simplicity, given the results, we believe that the proposals made in this section could be adopted.

On the other hand, both the Universal proposed for the SVO type and the version of the same Universal without considering alternative orders are positioned as Mid-Universals. In the first case close to the limit between being a Mid-Universal with high values and being a High-Universal with low values. In the second case with a high representativeness of the group to which it belongs. Therefore, we propose to discard any approximation similar to the one made here for the SVO type and focus on the results obtained for the SOV type, which we have been able to obtain interesting rules. We can summarise our findings as in table 21.12.

Summary of characteristics of the Universal 7 analysis:
1. The compliance of the original Universal is 89%.
2. Compliance in SVO languages without alternative basic order with adverbial modifiers preceding is 64%.
3. Compliance with the original Universal without the restriction of no alternative basic word order or only OSV is 85%.
4. Compliance with the Universal in SVO languages removing the restriction that there is no alternative basic word order is 53%.
5. Compliance with the original Universal with the adverb modifying any part of speech is 85%.
6. The fulfillment of the original Universal with the adverb modifying adjectives is 100%.
7. The Universal proposed by Greenberg is a High-Universal.
8. The formulation of a similar Universal for SVO languages is a Mid-Universal and, therefore, not desirable.
9. The fulfillment of the original Universal by specifying the modification of adjectives by adverbs gives clear and positive results.

Table 21.12: Summary of Characteristics of the Universal 7 Analysis

21.8 Universal 8

UNIVERSAL 8: *When a yes-no question is differentiated from the corresponding assertion by an intonational pattern, the distinctive intonational features of each of these patterns are reckoned from the end of the sentence rather than from the beginning.*

21.8.1 Current Limitation: Data no Available

This Universal cannot be checked using the systems and tools shown above. The formulation is based on “intonational features”, which indicate that they are placed at the end of the sentence in yes-no interrogatives that use this element to differentiate themselves from assertions. The Universal Dependencies project does not collect any information related to intonation or similar elements, which escape from syntax and some morphosyntactic items.

21.9 Universal 9

UNIVERSAL 9: *With well more than chance frequency, when question particles or affixes are specified in position by reference to the sentence as a whole, if initial, such elements are found in prepositional languages, and, if final, in postpositional.*

21.9.1 Current Limitation: Data no Available

This Universal cannot be verified due to the difficulty of obtaining information from different stated elements. First of all, in the implicational part, we already find an element impossible to check, in this case, because of the distinction between “by reference to the sentence as a whole” and “by reference to a particular word in the sentence” of the following Universal 10 (21.10). In this case, the limitation is produced by the syntactic analysis scheme on which UD is based, the Stanford dependencies, that is, the dependencies proposed by Tesnière (2015). In dependency analysis, any word depends on another word, which is its head. In this case, it is impossible to differentiate cases in which a word’s reference (or dependency) is to the utterance in general or only to the word on which it depends.

In addition, we must add the fact that it refers to “question particles or affixes”. About question particles, we have already commented in some Universal that there is some difficulty. However, it is more caused by annotation problems and not so much by the limitations of the tool. The information concerning whether a particle is interrogative or not is prepared in the proposed annotation scheme. However, some treebank annotators have not provided this information and limit themselves to providing the part of speech, severely limiting this and other possible studies.

Regarding affixes, we should point out that we do not have detailed information on their behavior in Universal Dependencies. The syntactic focus of the tool limits aspects related to the morphological realization of derivation and inflection, as may be the case of affixes. Therefore, not only do we have no information on when a language has interrogative affixes, but we do not have any specific information on language affixes in general.

21.10 Universal 10

UNIVERSAL 10: *Question particles or affixes, when specified in position by reference to a particular word in the sentence, almost always follow that word. Such particles do not occur in languages with dominant order VSO.*

21.10.1 Current Limitation: Data no Available

In the case of this Universal, we face the same problems as in Universal 9, (21.9). We have again the problem of the interpretation of the affixing of the interrogative element to the whole sentence or a particular word, the mislabeling of interrogative information in the particles of some treebanks, and the null information about affixes.

21.11 Universal 11

UNIVERSAL 11: *Inversion of statement order so that verb precedes subject only occurs in languages where the question word or phrase is normally initial. This same inversion occurs in yes-no questions only if it also occurs in interrogative word questions.*

21.11.1 Formalization Proposed for Grew-Match

The first we have to do regarding this Universal is to check the basic word order of each language. We have to note that it mentions a change from SV to VS. Then, we must discard all the languages that already have VS order. In order to do this, again, we have to reuse the patterns of Universal 1 (21.1) which we do not explain again because they are already widely used in this work. Then, we will only count SVO, SOV, and OSV languages.

The next step is checking if these languages tend to place the question word in the first position in the construction. A question word is usually known in English as W(H)-words. Also, we have to mention that it might not be the first word in the utterance since there can be other clauses; then we have to look for the first element in the affected part (or maybe the question word is preceded by prepositions, articles, etc.). We want to go further and analyze prototypical interrogative WH questions and every kind of structure that actually is a WH question. We also must mention that we are consciously ignoring languages that use another strategy for marking WH-questions as particles (PART) since this is checked separately in another Universal. If someone wants to check whether a language tends to put these elements at the beginning, we must formalize them like (21.97).

$$\begin{array}{l}
 \%11 - QW - initial \\
 pattern\{QW[upos = PRON, PronType = "Int"]; V[upos = \\
 VERB|AUX]; V- > QW; QW << V\}
 \end{array} \quad (21.97)$$

In (21.97), we are looking for question words in content questions (pronouns that are interrogative), and we isolate the interrogation phrase (in the case of an indirect question, for example) by asking for a pronoun modifying a verb dependency. Then, we ask for the order restriction. If we want to check the opposite order, the one not proposed by Greenberg, we should query (21.98).

$$\begin{array}{l}
\%11 - QW - final \\
pattern\{QW[upos = PRON, PronType = "Int"]; V[upos = \\
VERB|AUX]; V - > QW; QW >> V\}
\end{array} \tag{21.98}$$

We have to note that we cannot obtain results from all corpora, not only because some of them use particles but also because some languages use the same word for interrogations rather than for creating relative constructions. This is found with the particular label “PronType=Int,Rel”. In order to be more descriptive, we decided to add the query of this type of construction independently so that we can extend the claims of Greenberg not only to interrogations but also to relatives. The problem is that some treebanks (especially low-resourced languages that coincide with pronouns for content questions at the end) do not have a good annotation. For example, only “PRON” is tagged in Vietnamese, but not the “Int” feature. For this, we do not expect to find interesting results in this query. To obtain this, we should propose (21.99) and (21.100).

$$\begin{array}{l}
\%11 - QW - initial - INT - REL \\
pattern\{QW[upos = PRON, PronType = "Int, Rel"]; V[upos = \\
VERB|AUX]; V - > QW; QW << V\}
\end{array} \tag{21.99}$$

In (21.99), the interrogative or relative clause is placed after the verb.

$$\begin{array}{l}
\%11 - QW - final - INT - REL \\
pattern\{QW[upos = PRON, PronType = "Int, Rel"]; V[upos = \\
VERB|AUX]; V - > QW; QW >> V\}
\end{array} \tag{21.100}$$

In (21.100), the interrogative or relative clause comes before the verb.

When we query, for experimental purposes, the opposite situation, that is, where this interrogative pronoun is after the verb, we find two different trends:

If the language tends to have these word questions at the beginning, the cases in the last part (or at least after their governor) are always cases where this interrogative question is converted to the relative function.

The other possibility is that the language tends to place these word questions at the end. We do not have much information in this case due to the annotation errors. Thus, we cannot check with the same confidence what happens when an alternative is in the first position.

Once we have all these results, we can then search for the movement proposal of Greenberg. That is, to check if the subject is after the verb in these questions. Also, we must add one restriction regarding the equivalent of ‘who’ in any language (or ‘what’ sometimes, that is, anything that behaves as a subject) since the interrogative pronoun is functioning as a subject and then, there is a conflict between the subject before verb and the wh-word initial requirements. Anyway, we will check these situations independently to confirm our hypothesis: wh-question initial is a stronger requirement than the verb before the subject in all languages.

First, we query all the wh-questions except the ones in which the wh-word acts as a sentence subject. Moreover, we ask for the two possible orders (subject before verb and verb before subject). We check the possibility of placing the wh-question word before and the subject appearing before or after (21.101) and (21.102).

$$\begin{array}{l}
 \%11 - QWbef - SUBJaft \\
 pattern\{QW[upos = PRON, PronType = "Int"]; V[upos = \\
 VERB|AUX]; V - > QW; QW << V; S[upos = \\
 NOUN|PROPN|PRON]; V - [subj] - > S; V << S\} \\
 without\{QW[upos = PRON, PronType = "Int"]; V[upos = \\
 VERB|AUX]; V - [subj] - > QW; QW << V\}
 \end{array} \tag{21.101}$$

In (21.101), the interrogative element appears before the verb and the subject after the verb.

$$\begin{array}{l}
 \%11 - QWbef - SUBJbef \\
 pattern\{QW[upos = PRON, PronType = "Int"]; V[upos = \\
 VERB|AUX]; V - > QW; QW << V; S[upos = \\
 NOUN|PROPN|PRON]; V - [subj] - > S; V >> S\} \\
 without\{QW[upos = PRON, PronType = "Int"]; V[upos = \\
 VERB|AUX]; V - [subj] - > QW; QW << V\}
 \end{array} \tag{21.102}$$

In (21.102), the interrogative element appears before the verb and the subject before the verb.

We can also check what happens when the question word is placed at the end in (21.103) and (21.104).

$$\begin{array}{l}
\%11 - QW_{aft} - SUBJ_{aft} \\
pattern\{QW[upos = PRON, PronType = "Int"]; V[upos = \\
VERB|AUX]; V - > QW; QW >> V; S[upos = \\
NOUN|PROPN|PRON]; V - [subj] - > S; V << S\} \\
without\{QW[upos = PRON, PronType = "Int"]; V[upos = \\
VERB|AUX]; V - [subj] - > QW; QW << V\}
\end{array} \tag{21.103}$$

In (21.103), the interrogative element appears after the verb and the subject after the verb.

$$\begin{array}{l}
\%11 - QW_{aft} - SUBJ_{bef} \\
pattern\{QW[upos = PRON, PronType = "Int"]; V[upos = \\
VERB|AUX]; V - > QW; QW >> V; S[upos = \\
NOUN|PROPN|PRON]; V - [subj] - > S; V >> S\} \\
without\{QW[upos = PRON, PronType = "Int"]; V[upos = \\
VERB|AUX]; V - [subj] - > QW; QW << V\}
\end{array} \tag{21.104}$$

In (21.102), the interrogative element appears after the verb and the subject before the verb.

Then, we want to check what happens with the question words functioning as a subject. That is, to see if they behave like the rest of the question words (our hypothesis) or behave like the verb before subject restriction. For this, we propose (21.105) and (21.106).

$$\begin{array}{l}
\%11 - subjQW_{before} \\
pattern\{QW[upos = PRON, PronType = "Int"]; V[upos = \\
VERB|AUX]; V - [subj] - > QW; QW << V\}
\end{array} \tag{21.105}$$

In (21.105), the interrogative element is a subject that appears before the verb.

$$\begin{array}{l}
\%11 - subjQW_{after} \\
pattern\{QW[upos = PRON, PronType = "Int"]; V[upos = \\
VERB|AUX]; V - [subj] - > QW; QW >> V\}
\end{array} \tag{21.106}$$

In (21.106), the interrogative element is a subject that appears after the verb.

It is also interesting to check how ambiguous pronouns or interrogatives behave in the same situations regarding verb anticipation of the subject (or what happens with subject interrogative/relative pronouns). For this, we repeat the patterns but change interrogatives to this ambiguous kind as in (21.107) - (21.112).

$$\begin{array}{l}
\%11 - QWrelint - bef - SUBJaft \\
pattern\{QW[upos = PRON, PronType = "Int, Rel"]; V[upos = \\
VERB|AUX]; V - > QW; QW << V; S[upos = \\
NOUN|PROPN|PRON]; V - [subj] - > S; V << S\} \\
without\{QW[upos = PRON, PronType = "Int, Rel"]; V[upos = \\
VERB|AUX]; V - [subj] - > QW; QW << V\}
\end{array} \tag{21.107}$$

In (21.107), the interrogative or relative element appears before the verb and the subject after the verb.

$$\begin{array}{l}
\%11 - QWrelint - bef - SUBJbef \\
pattern\{QW[upos = PRON, PronType = "Int, Rel"]; V[upos = \\
VERB|AUX]; V - > QW; QW << V; S[upos = \\
NOUN|PROPN|PRON]; V - [subj] - > S; V >> S\} \\
without\{QW[upos = PRON, PronType = "Int, Rel"]; V[upos = \\
VERB|AUX]; V - [subj] - > QW; QW << V\}
\end{array} \tag{21.108}$$

In (21.108), the interrogative or relative element appears before the verb and the subject before the verb.

$$\begin{array}{l}
\%11 - QWrelint - aft - SUBJaft \\
pattern\{QW[upos = PRON, PronType = "Int, Rel"]; V[upos = \\
VERB|AUX]; V - > QW; QW >> V; S[upos = \\
NOUN|PROPN|PRON]; V - [subj] - > S; V << S\} \\
without\{QW[upos = PRON, PronType = "Int, Rel"]; V[upos = \\
VERB|AUX]; V - [subj] - > QW; QW << V\}
\end{array} \tag{21.109}$$

In (21.109), the interrogative or relative element appears after the verb and the subject after the verb.

$$\begin{array}{l}
\%11 - QWrelint - aft - SUBJbef \\
pattern\{QW[upos = PRON, PronType = "Int, Rel"]; V[upos = \\
VERB|AUX]; V - > QW; QW >> V; S[upos = \\
NOUN|PROPN|PRON]; V - [subj] - > S; V >> S\} \\
without\{QW[upos = PRON, PronType = "Int, Rel"]; V[upos = \\
VERB|AUX]; V - [subj] - > QW; QW << V\}
\end{array} \tag{21.110}$$

In (21.110), the interrogative or relative element appears after the verb and the subject before the verb.

$$\begin{array}{l}
\%11 - subjQWrelint - before \\
pattern\{QW[upos = PRON, PronType = "Int, Rel"]; V[upos = \\
VERB|AUX]; V - [subj] - > QW; QW << V\}
\end{array} \tag{21.111}$$

In (21.111), the interrogative or relative element is a subject that appears before the verb.

$$\begin{array}{l}
\%11 - subjQWrelint - after \\
pattern\{QW[upos = PRON, PronType = "Int, Rel"]; V[upos = \\
VERB|AUX]; V - [subj] - > QW; QW >> V\}
\end{array} \tag{21.112}$$

In (21.112), the interrogative or relative element is a subject that appears after the verb.

In the same way, we are tempted to check what happens with pure relative pronouns. We might want to know the behavior of the verb movement (or the subject before restriction). For this, we check, again, the same pattern but only checking relatives as in (21.113) - (21.118).

$$\begin{array}{l}
\%11 - RW - bef - SUBJaft \\
pattern\{RW[upos = PRON, PronType = "Rel"]; V[upos = \\
VERB|AUX]; V - > RW; RW << V; S[upos = \\
NOUN|PROPN|PRON]; V - [subj] - > S; V << S\} \\
without\{RW[upos = PRON, PronType = "Rel"]; V[upos = \\
VERB|AUX]; V - [subj] - > RW; RW << V\}
\end{array} \tag{21.113}$$

In (21.113), the relative element appears before the verb and the subject after the verb.

$$\begin{array}{l}
\%11 - RW - bef - SUBJbef \\
pattern\{RW[upos = PRON, PronType = "Rel"]; V[upos = \\
VERB|AUX]; V - > RW; RW << V; S[upos = \\
NOUN|PROPN|PRON]; V - [subj] - > S; V >> S\} \\
without\{RW[upos = PRON, PronType = "Rel"]; V[upos = \\
VERB|AUX]; V - [subj] - > RW; RW << V\}
\end{array} \tag{21.114}$$

In (21.114), the relative element appears before the verb and the subject before the verb.

$$\begin{array}{l}
\%11 - RW - aft - SUBJaft \\
pattern\{RW[upos = PRON, PronType = "Rel"]; V[upos = \\
VERB|AUX]; V - > RW; RW >> V; S[upos = \\
NOUN|PROPN|PRON]; V - [subj] - > S; V << S\} \\
without\{RW[upos = PRON, PronType = "Rel"]; V[upos = \\
VERB|AUX]; V - [subj] - > RW; RW << V\}
\end{array} \tag{21.115}$$

In (21.115), the relative element appears after the verb and the subject after the verb.

$$\begin{array}{l}
\%11 - RW - aft - SUBJbef \\
pattern\{RW[upos = PRON, PronType = "Rel"]; V[upos = \\
VERB|AUX]; V - > RW; RW >> V; S[upos = \\
NOUN|PROPN|PRON]; V - [subj] - > S; V >> S\} \\
without\{RW[upos = PRON, PronType = "Rel"]; V[upos = \\
VERB|AUX]; V - [subj] - > RW; RW << V\}
\end{array} \tag{21.116}$$

In (21.116), the relative element appears after the verb and the subject before the verb.

$$\begin{array}{l}
\%11 - subjRW - before \\
pattern\{RW[upos = PRON, PronType = "Rel"]; V[upos = \\
VERB|AUX]; V - [subj] - > RW; RW << V\}
\end{array} \tag{21.117}$$

In (21.117), the relative element is a subject that appears before the verb.

$$\begin{array}{l}
\%11 - subjRW - after \\
pattern\{RW[upos = PRON, PronType = "Rel"]; V[upos = \\
VERB|AUX]; V - [subj] - > RW; RW >> V\}
\end{array} \tag{21.118}$$

In (21.118), the relative element is a subject that appears after the verb.

So far, we have been solving the central part of the Universal proposed by Greenberg (and checking other possible Universals, such as the same application to relative pronouns). However, this Universal has an additional Universal, expressed in the last sentence, related to yes-no questions. Unfortunately, we cannot analyze this part with the available data. First of all, many Treebanks do not have the information of the particles being interrogative, although the tag exists. Secondly, the possibility of including a specific tag for yes/no questions, which would facilitate this analysis, has not been considered.

We believe that in the future, it would be easier to analyze this type of Universals if we worked with interglossae. This helpful element in Linguistic Typology should be extended to Treebanks. However, another option is to add a label to reflect this type of information as a yes/no question. In no case should it imply a reformulation of the present label system or the already annotated treebanks. Therefore, as we consider it as additional information, we propose to do it as an additional tag, headed by the “@” element, which is used to add other additional information, which can be added or not, without affecting the rest of the information at all. Something along the lines of “Int@PolarQ”.

We should note that after forming this proposal, we have investigated whether this concern is shared by more expert researchers in the field. We noticed comments in the Universal dependencies annotation guides where researchers also raise this problem. However, no consensus has been reached. We believe that the main problem is that it is impossible to add punctual patches with a specific element to a scheme such as UD and that what should be done is to propose a comprehensive update with different tags that have been raised.

21.11.2 Trichotomous Expression Based on Grew-Match Data

As we have already commented in the formalization section, there are a series of limitations that do not allow the verification of the Universal as conceived by Greenberg. In this case, we can check the order of the interrogative particle in 40 languages. The total number of languages that can be analyzed falls below the potential total for two main reasons. First, we detected annotation errors in some of the Treebanks, in which the interrogative information is not marked in the interrogative elements, as we have observed in Turkish, for example. In such a language, we have such particles in the final position (“nerede”, for example), and no such morphological information appears in their multiple occurrences in the corpus annotations. On the other hand, we must remember that there

are some languages for which we do not have information because they are of the agglutinative type and are not marked correctly in Universal Dependencies. This resource does not include information on affixes, making it challenging to work with languages of this type.

However, of the 40 languages that we can analyze, we have determined that 93% of them exhibit the interrogative element in clause-initial position. This data would be represented in a trichotomous expression as in figure 21.23.

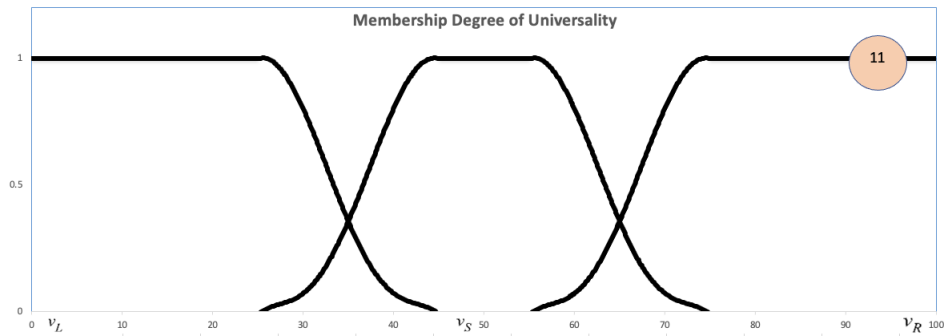


Figure 21.23: Trichotomous Expression for Universal 11

This new Universal that we have formulated or simplified (given the tool’s limitations) is a prototypical High-Universal. This makes it a rule to be considered from a typological point of view. However, we believe that a more varied, richer, and more representative selection and a better annotation of all types of languages will allow us to know the claim formulated here in a more reliable way, which we understand as a mere starting point to be confirmed or refuted. We can summarise our findings as in table 21.13.

Summary of characteristics of the Universal 11 analysis:
1. The UD annotation scheme makes it difficult to analyze this Universal in its original terms because of the influence of relative structures.
2. There are also difficulties in checking the Universal completeness because of UD’s lack of annotated data concerning yes/no questions.
3. Most languages place the interrogative element in the initial position (93%).
4. This unrestricted rule would be a High-Universal.

Table 21.13: Summary of Characteristics of the Universal 11 Analysis

21.12 Universal 12

UNIVERSAL 12: *If a language has dominant order VSO in declarative sentences, it always puts interrogative words or phrases first in interrogative word questions; if it has dominant order SOV in declarative sentences, there is never such an invariant rule.*

21.12.1 Formalization Proposed for Grew-Match

For the first part of the Universal, we must, again, decide each language’s dominant basic word order. In other words, we must query all the possible basic word orders as stated in Universal 1 (21.1).

$$\begin{array}{l}
 \%12 - SVO \\
 pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = \\
 PROPEN|NOUN]; V - [comp : obj] - > O; O[upos = \\
 PROPEN|NOUN]; V >> S; S << O; V << O\} \\
 without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
 \end{array} \tag{21.119}$$

In (21.119), to capture the SVO word order, we ask the subject to be before the verb and the object to be after the verb.

$$\begin{array}{l}
 \%12 - SOV \\
 pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = \\
 PROPEN|NOUN]; V - [comp : obj] - > O; O[upos = \\
 PROPEN|NOUN]; V >> S; S << O; V >> O\} \\
 without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
 \end{array} \tag{21.120}$$

In (21.120), to capture the SOV word order, we ask the subject to be before the verb and the object to be before the verb but not before the subject.

$$\begin{array}{l}
 \%12 - VSO \\
 pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = \\
 PROPEN|NOUN]; V - [comp : obj] - > O; O[upos = \\
 PROPEN|NOUN]; V << S; S << O\} \\
 without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
 \end{array} \tag{21.121}$$

In the VSO case, in (21.121), we want a verb before the subject and before the object, but the subject before the object.

$$\begin{aligned}
 &\%12 - VOS \\
 &pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = \\
 &PROPN|NOUN]; V - [comp : obj] - > O; O[upos = \\
 &PROPN|NOUN]; V << S; S >> O; V << O\} \\
 &without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
 \end{aligned} \tag{21.122}$$

In VOS word order, in (21.122), we formalize the restriction of verb being before the subject and before the object, but the object appearing before the subject.

$$\begin{aligned}
 &\%12 - OSV \\
 &pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = \\
 &PROPN|NOUN]; V - [comp : obj] - > O; O[upos = \\
 &PROPN|NOUN]; V >> S; S >> O; V >> O\} \\
 &without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
 \end{aligned} \tag{21.123}$$

In the OSV order, in (21.123), we ask for objects appearing before subjects and verbs, but subjects before verbs.

$$\begin{aligned}
 &\%12 - OVS \\
 &pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = \\
 &PROPN|NOUN]; V - [comp : obj] - > O; O[upos = \\
 &PROPN|NOUN]; V << S; S >> O; V >> O\} \\
 &without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
 \end{aligned} \tag{21.124}$$

Furthermore, in OVS order, in (21.124), we ask for objects appearing before subjects and verbs, but verbs before subjects.

Then, we have to calculate the dominance, as we mentioned, again. Finally, we have just to analyze languages that exhibit VSO basic word order for one part of the Universal and select SOV languages for the other part. Although we will also check SOV languages because they are pointed out by the Universal, we believe Greenberg's formalization is useless. In the case of this linguistic type, he states, "There is no such invariant rule". That is, the Universal applied to SOV is that there is no Universal tendency. Therefore, the predictive power is very low.

Then, we must reuse patterns already stated in Universal 11. Precisely, we query formalizations (21.125) and (21.126). Thanks to these checks, we will be able to determine whether these words always appear in the initial position or not.

$$\begin{array}{l} \%12 - QW - initial \\ pattern\{QW[upos = PRON, PronType = "Int"]; V[upos = \\ VERB|AUX]; V - > QW; QW << V\} \end{array} \quad (21.125)$$

In (21.125), we capture interrogative elements before the head verb.

$$\begin{array}{l} \%12 - QW - final \\ pattern\{QW[upos = PRON, PronType = "Int"]; V[upos = \\ VERB|AUX]; V - > QW; QW >> V\} \end{array} \quad (21.126)$$

In (21.126), we capture interrogative elements after the head verb.

Similarly, it is interesting to see if this behavior is extensible to relative pronouns. Therefore, we propose the formalizations (21.127) and (21.128). In these, we just check if the position of the relative element is always initial, in agreement with what occurs in interrogative words.

$$\begin{array}{l} \%12 - RW - initial \\ pattern\{RW[upos = PRON, PronType = "Rel"]; V[upos = \\ VERB|AUX]; V - > RW; RW << V\} \end{array} \quad (21.127)$$

In (21.127), we capture relative elements before the head verb.

$$\begin{array}{l} \%12 - RW - final \\ pattern\{RW[upos = PRON, PronType = "Rel"]; V[upos = \\ VERB|AUX]; V - > RW; RW >> V\} \end{array} \quad (21.128)$$

In (21.127), we capture relative elements after the head verb.

21.12.2 Limitations Based on Grew-Match Data

In this example, as in Universal 11, we find starting limitations in its formalization due to the unavailability of such information in the Universal Dependencies annotation scheme. In addition, we must also apply the restrictions on annotation errors discussed in the previous subsection here.

In this case, we must consider that it is Universal and applicable only to languages of type VSO, which, as we know, are very few (both in our set and the totality of languages of the world). Therefore, we believe that the sum of limitations and restrictions makes the number of testable languages very small. This ultimately makes the data set unworthy of representation because it is not illustrative.

21.13 Universal 13

UNIVERSAL 13: *If the nominal object always precedes the verb, then verb forms subordinate to the main verb also precede it.*

21.13.1 How to Check It in Typometrics

Technical Decisions:

1. Firstly, the Universal is analyzed with Typometrics (Gerdes et al., 2021).
2. We select the SUD annotation scheme because we can get results more comparable to the classical typological claims.
3. We select Graph 2D.
4. We set a filter of utterances to avoid misleading data.
5. We select *direction – cfc* in both axes.
6. The x-Axis can be *VERB – comp : obj – NOUN* (verb as head, noun as dependent, object as dependency), and the y-Axis can be *VERB – comp – VERB* (verb as head, another verb as dependent, which is an argument of the head verb).

In this case, since the Universal just mentions two individual relations, there is no need for further checking.

Table 21.14: Technical Decisions for Checking Universal 13 in Typometrics

21.13.2 Formalization Proposed for Grew-Match

If we want to check this same Universal via Grew-Match, we have to propose two patterns already pointed out in Universal 1, (21.129) and (21.130), and two new formalizations, (21.131) and

(21.132).

$$\begin{array}{l} \%13 - OBJ - bef \\ pattern\{V[upos = VERB]; V - [comp : obj] - > O; O[upos = \\ NOUN|PROPN]; V >> O\} \end{array} \quad (21.129)$$

In (21.129), we can see how we simply name V any POS that is $VERB$ and O any POS that is $NOUN$ or $PROPN$. We require O to be the object of V head and O to precede V .

$$\begin{array}{l} \%13 - OBJ - aft \\ pattern\{V[upos = VERB]; V - [comp : obj] - > O; O[upos = \\ NOUN|PROPN]; V << O\} \end{array} \quad (21.130)$$

In the case of (21.130), we repeat the same pattern but change the order sense and ask that V precedes O , i.e., “ $V << O$ ”.

$$\begin{array}{l} \%13 - VERBdependent - before \\ pattern\{V[upos = VERB]; V - [1 = comp] - > VD; VD[upos = \\ VERB]; V >> VD\} \end{array} \quad (21.131)$$

In (21.131), we repeat the naming of V and assign the VD tag to a POS of $VERB$. We ask V to be the head of VD , which acts as a dependent. The order we set is $V >> VD$; the dependent verb appears in the first position.

$$\begin{array}{l} \%13 - VERBdependent - after \\ pattern\{V[upos = VERB]; V - [1 = comp] - > VD; VD[upos = \\ VERB]; V << VD\} \end{array} \quad (21.132)$$

In (21.132), we repeat the pattern, but this time asking for $V << VD$, the verb as head appears in the first position.

Another option we have, thanks to corpus linguistics and Computational Linguistics, is what happens when the two parts of the Universal appear in a single utterance and to count the cases. What we want to check is whether the possible flexibility that may appear in languages concerning the presence of both orders ($VD << V$ and $V << VD$, for example), even if one of them predominates, is reduced in cases of simultaneous occurrence with the specific order OV or VO.

Confirmation of this hypothesis would induce a possible new Universal that would be encompassed within the recent studies that exist about syntactic harmony, which, in reality, is nothing more than the manifestation of the search for ease of processing (Futrell et al., 2020). Although it is more complicated to find sentences of this type in languages and, therefore, the frequency of occurrence is scarce, it can be a good starting point for similar future research as a proof of concept. For this purpose, we propose the patterns (21.133), (21.134), (21.135) and (21.136) patterns, which correspond to the different possibilities that exist:

$$\begin{array}{l}
 \%13 - OBJ - bef - VD - aft \\
 pattern\{V[upos = VERB]; O[upos = NOUN|PROPN]; V - [comp : \\
 obj] - > O; VD[upos = VERB]; V - [1 = comp] - > VD; V << \\
 VD; V >> O\}
 \end{array} \tag{21.133}$$

In (21.133), we can see how what we have done is to merge the previous patterns. We ask the previous V to have two arguments simultaneously, one O and the other VD . The labels O and VD still correspond to the previous values, as does V . In this case, we ask that the order be “ $V << VD; V >> O$ ”, that is, that the object appears before the head verb and that the dependent verb appears after the head verb.

$$\begin{array}{l}
 \%13 - OBJ - bef - VD - bef \\
 pattern\{V[upos = VERB]; O[upos = NOUN|PROPN]; V - [comp : \\
 obj] - > O; VD[upos = VERB]; V - [1 = comp] - > VD; V >> \\
 VD; V >> O\}
 \end{array} \tag{21.134}$$

In (21.134), we repeat the same pattern, but in this case, the order is the object and dependent verb before the head verb ($V >> VD; V >> O$).

$$\begin{array}{l}
 \%13 - OBJ - aft - VD - bef \\
 pattern\{V[upos = VERB]; O[upos = NOUN|PROPN]; V - [comp : \\
 obj] - > O; VD[upos = VERB]; V - [1 = comp] - > VD; V >> \\
 VD; V << O\}
 \end{array} \tag{21.135}$$

In (21.135), we propose the order $V >> VD; V <<< O$ for the structure, i.e., the object is placed after the head verb and the dependent verb before the head verb.

$$\begin{array}{l}
\%13 - OBJ - aft - VD - aft \\
pattern\{V[upos = VERB]; O[upos = NOUN|PROPN]; V - [comp : \\
obj]- > O; VD[upos = VERB]; V - [1 = comp]- > VD; V << \\
VD; V << O\}
\end{array} \tag{21.136}$$

Finally, in (21.136), we ask for the order $V << VD; V << O$ corresponding to both the dependent verb and the object appearing after the head verb.

Moreover, if we want, we can go further still and provide more precise and detailed results (although by further refining the level of detail, we will also find fewer occurrences to work with; that is why it is simply an experimental and novel proposal and the data should be extended in the future and contrasted). In other words, we propose a new, stricter pattern with the word order we require to determine the position of O and VD when they appear in the same direction concerning the verb head. Thus, we can establish a linguistic type of languages that tend to place the object closer to the verb or the dependent verb closer to the verb when they co-occur. Then, we can also compare these data with languages that tend to place one on either side, to see which direction is preferred. However, these types of decisions in languages are very complex, as different factors influence the choice. For example, a candidate to be the trigger would be a “short before long” constraint, which tries to place the shorter elements closer to the head to facilitate processing. In short, the patterns we propose are (21.137) - (21.140).

$$\begin{array}{l}
\%13 - OBJ - bef - VD - bef - OBJbef \\
pattern\{V[upos = VERB]; O[upos = NOUN|PROPN]; V - [comp : \\
obj]- > O; VD[upos = VERB]; V - [1 = comp]- > VD; V >> \\
VD; V >> O; VD >> O\}
\end{array} \tag{21.137}$$

(21.137) captures the dependent verbs preceding the main verb, the direct objects preceding the main verb, and the dependent verb in cooccurrence.

$$\begin{array}{l}
\%13 - OBJ - bef - VD - bef - OBJaft \\
pattern\{V[upos = VERB]; O[upos = NOUN|PROPN]; V - [comp : \\
obj]- > O; VD[upos = VERB]; V - [1 = comp]- > VD; V >> \\
VD; V >> O; VD << O\}
\end{array} \tag{21.138}$$

(21.138) captures dependent verbs preceding the main verb, direct objects preceding the main verb, and dependent verbs preceding objects in cooccurrence.

$$\begin{array}{l}
\%13 - OBJ - aft - VD - aft - OBJbef \\
pattern\{V[upos = VERB]; O[upos = NOUN|PROPN]; V - [comp : \\
obj] - > O; VD[upos = VERB]; V - [1 = comp] - > VD; V << \\
VD; V << O; VD >> O\}
\end{array} \tag{21.139}$$

(21.139) captures after dependent verbs of the main verb, direct objects after to the main verb, and preceding dependent verbs in cooccurrence.

$$\begin{array}{l}
\%13 - OBJ - aft - VD - aft - OBJaft \\
pattern\{V[upos = VERB]; O[upos = NOUN|PROPN]; V - [comp : \\
obj] - > O; VD[upos = VERB]; V - [1 = comp] - > VD; V << \\
VD; V << O; VD << O\}
\end{array} \tag{21.140}$$

(21.140) captures after dependent verbs of the main verb, direct objects after to the main verb, and dependent verbs in cooccurrence.

In the patterns, the only thing we have done is, as mentioned above, to add an extra order constraint in the case of occurrence in the same direction. In the case of occurrence on different sides, adding this restriction is either redundancy or the formalization of something impossible logically. Therefore, we do not include the patterns (21.141) and (21.142) (because they are redundant and already have these data in the patterns with a lower precision) (21.143) and (21.144) (because they are not possible).

$$\begin{array}{l}
\%13 - OBJ - bef - VD - aft - OBJbef \\
pattern\{V[upos = VERB]; O[upos = NOUN|PROPN]; V - [comp : \\
obj] - > O; VD[upos = VERB]; V - [1 = comp] - > VD; V << \\
VD; V >> O; VD >> O\}
\end{array} \tag{21.141}$$

(21.141) captures after dependent verbs of the main verb, direct objects before the main verb, and dependent verbs in cooccurrence.

$$\begin{array}{l}
\%13 - OBJ - bef - VD - aft - OBJaft \\
pattern\{V[upos = VERB]; O[upos = NOUN|PROPN]; V - [comp : \\
obj] - > O; VD[upos = VERB]; V - [1 = comp] - > VD; V << \\
VD; V >> O; VD << O\}
\end{array} \tag{21.142}$$

(21.142) captures after dependent verbs of the main verb, direct objects preceding the main verb,

and after the dependent verbs in cooccurrence.

$$\begin{array}{l}
 \%13 - OBJ - aft - VD - bef - OBJbef \\
 pattern\{V[upos = VERB]; O[upos = NOUN|PROPN]; V - [comp : \\
 obj] - > O; VD[upos = VERB]; V - [1 = comp] - > VD; V >> \\
 VD; V << O; VD >> O\}
 \end{array} \tag{21.143}$$

(21.143) captures after dependent verbs of the main verb, direct objects after to the main verb, and preceding to dependent verbs in cooccurrence.

$$\begin{array}{l}
 \%13 - OBJ - aft - VD - bef - OBJaft \\
 pattern\{V[upos = VERB]; O[upos = NOUN|PROPN]; V - [comp : \\
 obj] - > O; VD[upos = VERB]; V - [1 = comp] - > VD; V >> \\
 VD; V << O; VD << O\}
 \end{array} \tag{21.144}$$

(21.144) captures after dependent verbs of the main verb, direct objects after to the main verb, and dependent verbs in cooccurrence.

Another possibility is to check the distance (and not the direction) in the relationship between these elements. In this case, we will not present total results since this objective is far from the object of study of the thesis, and we would have to deal in depth with aspects such as those mentioned above that condition the dependency length. We briefly show how, methodologically, a linguist obtains such data thanks to Grew-Match.

We propose to take advantage of the pattern previously created and shown, for example, (21.137), and add in the first dependency relation, that of the object, before the head: “*o* :”. Then, in the dependent verb’s second dependency, add “*v* : before the head”. Subsequently, we must activate the option to do clustering by key and enable two different clustering that allow us to create a table with two axes. In the first clustering, we put “*o.length*” and in the second, “*v.length*”. This will help us quickly and visually compare the different distance length correspondences between the two dependents of the head verb. In the general example we have cited, we have chosen to eliminate any direction restriction so that the distance calculated is absolute. If a researcher wanted to calculate the distance given a specific constraint, it would suffice to add the desired direction for the distance to be analyzed using “<<” as we have repeatedly shown in this paper. The proposed pattern, therefore, would look like in (21.145).

v.length		o.length		v.length		o.length		v.length		o.length		v.length		o.length		v.length		o.length	
47	1	8	3	6	4	5	2	5	6	4	9	2	14	2	8	12	—	*	—
39	3	33		1				1		1				1				2	
34	2	6	8	4				4		2		1		1				8	
8	4	6				1												1	
4	1					4													
2	5											1						1	
1	11									1									
1	19	1																	
1	6				1														
1	7	1																	

Figure 21.24: Distance of O and VD to V Head in Spanish (AnCora)

v.length o.length		245 2	182 3	64 4	23 5	22 1	18 6	6 7	1 12	1 8
393 1	245	108	23	8	1	7				1
122 2		74	32	7	3	4	2			
28 3			9	4	10	3	2			
13 4				4	5	4				
2 5					1		1			
2 6					1		1			
1 10									1	
1 8					1					

Figure 21.25: Distance of O and VD to V Head in Classical Chinese (Kyoto)

%13 – OBJ – VD – distance

pattern{V[upos = VERB]; O[upos = NOUN|PROPN]; o : V – [comp :
obj] – > O; VD[upos = VERB]; v : V – [1 = comp] – > VD; V >>
VD; V >> O; VD >> O}

(21.145)

(21.145) is a pattern modified to show the distance of the dependent verbs preceding the main verb, the direct object after the main verb, and the dependent verb preceding the object in cooccurrence.

To exemplify the results obtained, we can show the tables we have created from a Spanish corpus in figure 21.24 and a Classical Chinese corpus in figure 21.25.

As we can see in 21.24, the most common distance in Spanish is 1 for the verb dependent on another verb as head and 3 for the nominal object when they occur together. The data for the following positions in the ranking support the theory that the verb always has priority in Spanish

to appear before the object when dependent on the same verb in the sentence. This can be seen in only four occurrences in which the nominal object has a distance of 1. In contrast, in 21.25, in Classical Chinese, clearly most frequent relationship is the placement of the nominal object with a distance of 1 and the dependent verb with a distance of 2. The following most frequent distance is of the object at a distance of 1 and the verb at a distance of 3. Furthermore, when the object is at a distance of 2, the verb is at a distance of 3, not being at any time 1 or 2, i.e., prior to the nominal object. This restriction seems very strong. In other words, Spanish seems to prefer structures equivalent to “advance the process” (lit. order), while Classical Chinese would prefer the equivalent of “make the process advance” (lit. order).

These results, once analyzed, could be compared with those of the distance of these same elements but not occurring in the same sentence. These data can be quickly obtained in Typometrics by creating a 2D graph, with $VERB - comp : obj - NOUN$ in x-Axis and $VERB - comp - VERB$ in y-Axis. This time, the “measure” to select is *distance - cfc* in both cases.

21.13.3 Graphs from Typometrics

The verification of this Universal implication is quite clear using Typometrics given its simplicity, as shown in figures 21.26 and 21.27.

The implication that restricts which languages the Universal affects invites us to review the left-hand side of the graph, which corresponds to the OV languages (nominally). If Greenberg’s proposal is accurate, these languages on the left-hand side of the figure 21.26 will all appear in the lower corner, in harmony with word order (both elements placed before each other). As can be seen, languages exist in both corners (and from different language families). Therefore, it does not seem that the Universal is well formulated.

Although we cannot identify any pattern with a particular shape in this Universal, we can identify a resistance of languages to be placed in the opposite corner, as shown in the figure 21.27. Since we are on the right-hand side of the graph, we are referring to nominally OV languages. In this case, the word order has a clear harmony, and the verbs dependent on a head verb are also postponed. In the blue square, we show no languages, as they all appear in the upper right-hand corner. Therefore, we believe it would be more productive to discard the Universal proposed by Greenberg and follow this new proposal instead. We propose revising the quantitative data via Grew-Match to check this more precisely.

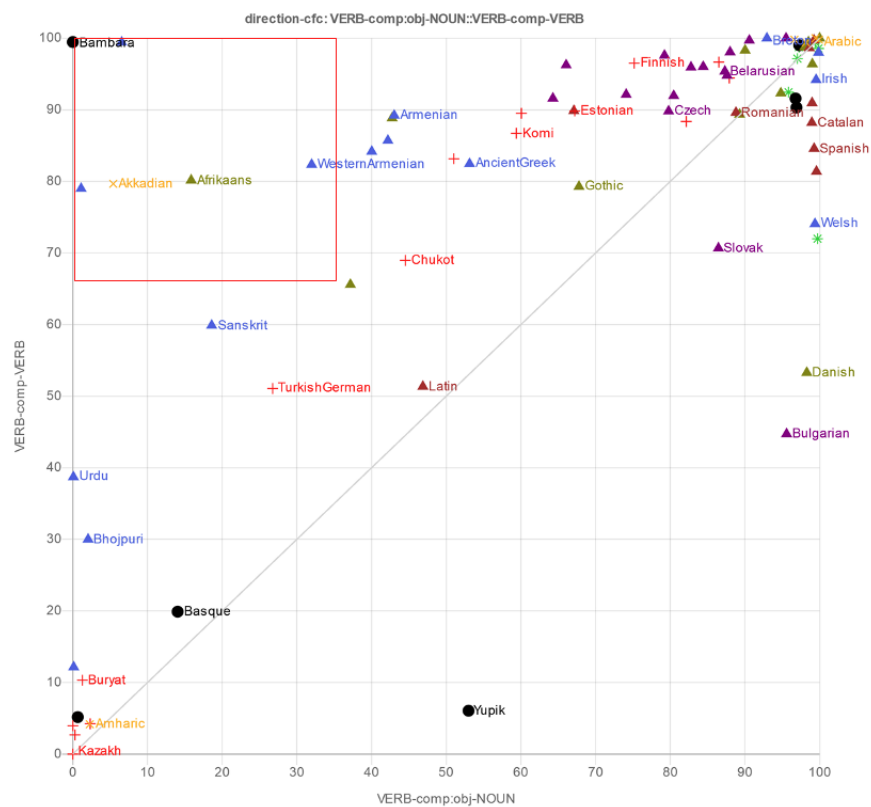


Figure 21.26: Nominal Direct Objects and Verbal Direct Objects Direction Checking

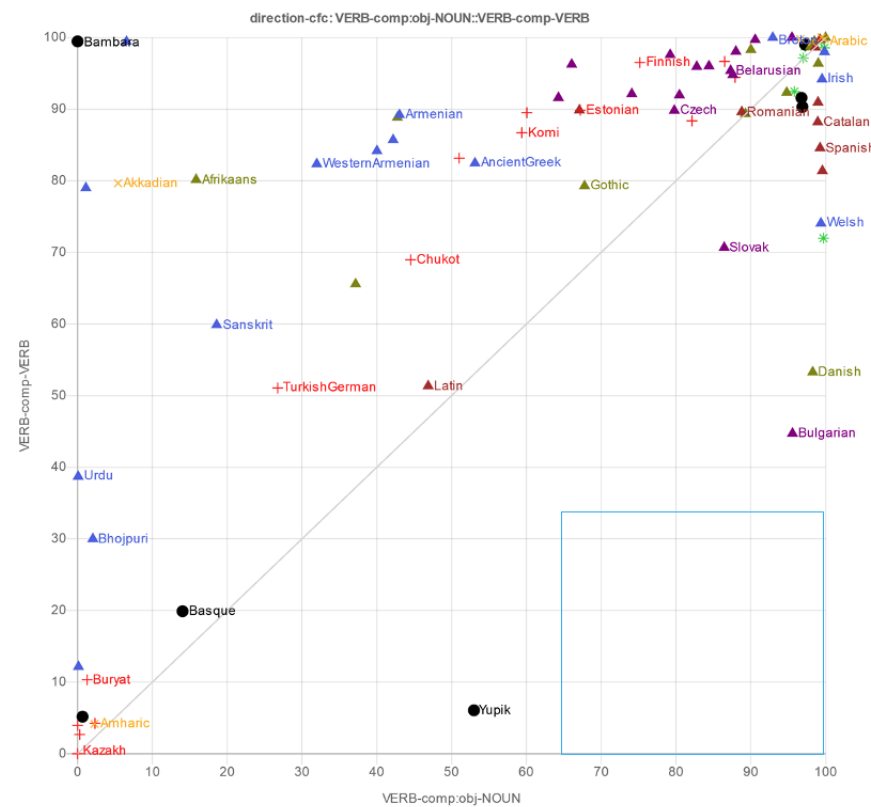


Figure 21.27: Nominal Direct Objects and Verbal Direct Objects Direction Proposal

21.13.4 Trichotomous Expression Based on Grew-Match Data

From this approach, we obtain that the Universal proposed by Greenberg is fulfilled in 57% of the cases, far from what would be desirable to be able to formulate a Universal. On the other hand, the High-Universal we propose is fulfilled in 97% of cases. Moreover, thanks to the possibility of creating more complex patterns, we can check the direction of both elements in cooccurrence. There are three possibilities:

1. First, we may obtain the same results, meaning that such cooccurrence does not imply any extra constraints.
2. Secondly, we may obtain higher results relative to the same direction in cooccurrence, which would support a functional justification of order harmony.
3. Third, we may obtain lower results relative to the same direction in cooccurrence, which would support a functional justification for the closest possible location between dependents and heads.

In this case, we obtain those languages whose order is OV and which also have dependent verbs preceding the verb 60% of the time in the same example. About the languages whose order is OV and which are also present in this example dependent verbs after the verb, we obtain 88% of compliance. The difference in the first case does not seem significant. On the other hand, in the second case, there is a slight tendency to favor the third situation described. In the case of two competing functional solutions that facilitate processing (same direction or favoring proximity to the head), it seems that a smaller distance between dependents is chosen. However, we should point out that the languages contributing to this slight decrease in direction consistency are part of the Sino-Tibetan family (except Classical Chinese). Therefore, we believe it is necessary to review these data in greater depth when more languages become available to help us confirm whether this is a trend of some languages or, indeed, the atypical (or Low-Universal) characteristic of a particular language family or area.

If we want to transfer this data to the trichotomous model, we obtain a result like the one shown in figure 21.28.

It can be seen that Greenberg's original proposal is a Mid-Universal and, therefore, not a relevant rule to explain the behavior of OV languages. Thus, it is understood that the location of other-dependent verbs in OV languages is arbitrary or, at least, there is no correlation between

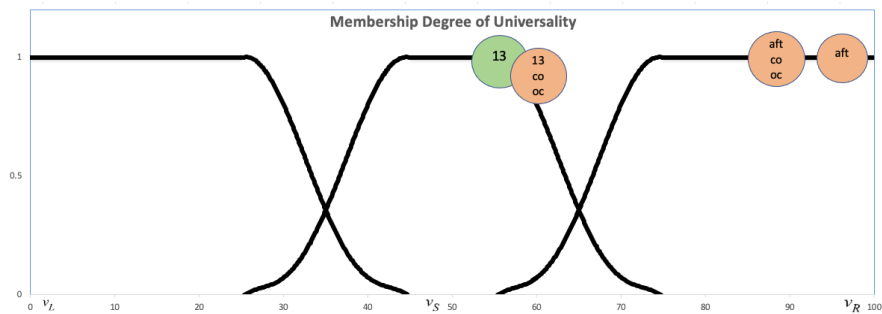


Figure 21.28: Trichotomous Expression for Universal 13

these two features. The predictive implication must be changed. At this same place in the model is the variation where both parameters appear in the same occurrence and the conclusions are the same. The position in which they are placed within the set (representative of a Mid-Universal) leaves no doubt that they are part of it.

The Universal we propose is a High-Universal and representative of its set. So, is the possible Universal that takes into account simultaneous occurrence? Therefore, in this case, we propose to discard Greenberg’s original Universal and, instead, encourage the adoption of our proposal, which offers a much clearer and stronger correlation. If a language has a nominal VO order, it also places verbs dependent on another verb after it. We can summarise our findings as in table 21.15.

Summary of characteristics of the Universal 13 analysis:
1. Typometrics: Languages tend to place the nominal direct object harmoniously and the dependent verb after (upper right corner).
2. Typometrics: There are languages such as Afrikaans, Akkadian, or Bambara, which are exceptions to Greenberg’s Universal (top left corner).
3. Typometrics: No languages with dominant nominal VO order tend to place the dependent verb before the main verb (bottom right corner).
4. The original Universal has a 57% compliance rate.
5. The Universal we propose concerning languages with a dominant nominal VO order, which tends to place the dependent verb before the main verb, also after has a 97% compliance.
6. The result of the analysis of these two Universals in contexts of cooccurrence is practically the same; no changes can be seen.
7. Greenberg’s original Universal is a Mid-Universal.
8. The Universal we propose is a High-Universal.
9. The Sino-Tibetan family shows peculiar behavior in the cooccurrences.

Table 21.15: Summary of Characteristics of the Universal 13 Analysis

21.14 Universal 14

UNIVERSAL 14: *In conditional statements, the conditional clause precedes the conclusion as the normal order in all languages.*

21.14.1 Checking Combining PanLex and Grew-Match

In order to work with this very specific Universal proposed by Greenberg, we must propose a novel and more complex methodology. The first challenge is determining how to capture and isolate a conditional structure from what Greenberg calls a “conclusion”. To do this, we must create a pattern containing three elements in UD. The first, labeled “*IF*”, is a *SCONJ* (although in a given language, it is not strictly a subordinating conjunction, it is labeled so in UD so that we can compare languages with each other). It should be noted that some languages do not use a word to generate the conditional (always labeled as *SCONJ*) but use some element attached to another word. In that case, UD labels these cases differently and unsystematically, often *PART* particles. The main problem is that in UD, there is no morphological information about the different constituents of a word. We have even thought of an alternative way to capture the cases in which it is not expressed by a word, as can be seen in the proposed formalization in (21.146). In this case, we look for the particle or desinence that indicates ‘conditional’ in a language such as Basque. Moreover, we carry out an additional operation in which we propose to contrast the word’s form with its lemma formally. If the form and the lemma contain the element we are looking for (for example, a verb whose root begins with “ba” in Basque), it is discarded since it would not correspond to this process. However, in the case of this same language, such a prefix is also used as an emphatic affirmative and for other uses. Since we do not have complete information on all this in Universal Dependencies, it is impossible to know when it is conditional in the different languages. Therefore, if a language has such a conditional mechanism, it cannot be treated at present, given the limitations of UD, and will have to be treated in the future once we have this data.

%14 – <i>IF</i> – <i>ba</i> – <i>part</i> <i>pattern</i>{<i>GOV</i> – [<i>advcl</i>] – > <i>DEP</i>; <i>DEP</i>[<i>upos</i> = <i>VERB</i>, <i>form</i> = <i>re</i>“<i>ba</i>. * ”, <i>lemma</i> <> <i>re</i>“<i>ba</i>. * ”] }	(21.146)

This *IF* element has to have one more characteristic, besides being *SCONJ*: to have the lemma (or lemmas) specific to each language with the meaning of conditional, which, for example, in Spanish would be “si” and which, therefore, we must fill in with the lemma of each language and formalize as *lemma* = “*x*”. Subsequently, we name an element as *VIF*, simply a verb (*VERB*) or an auxiliary (*AUX*) to which we will attribute the value of being the verb of the *IF* structure. Finally, we propose the element *VMAIN*, another verb or auxiliary (*VERB* or *AUX*) corresponding to the verb of the conclusion structure.

Subsequently, we require that there is a dependency relation between *VMAIN* and *VIF*, where the former is a head, and the latter is dependent, in *advcl* relation, since *IF* structures always have this type of structure in UD. In addition, we add the dependency between *VIF* and *IF*, where the former is a head and the latter is a dependent. There is a *mark* relation since in UD, the *SCONJ* of this type is always considered a mark type to the verb that makes it a specific type, a relation comparable to that affecting an *ADP* dependent on a *NOUN*.

Finally, to indicate that the conditional structure is prior to the conclusion, we add the constraint *IF* <<< *VMAIN*. It is also possible to add the constraint *VIF* <<< *VMAIN*, but this does not influence the results at all since we have not found any language in which the *VMAIN* appears between *IF* and *VIF*, so you can leave a single order constraint if you want to simplify the pattern. In summary, the two patterns we have developed are (21.147) and (21.148):

<pre>%14 – IF – before pattern{IF[upos = SCONJ, lemma = “X”]; VIF[upos = VERB AUX]; VMAIN[upos = VERB AUX]; VIF – [mark] – > IF; VMAIN – [advcl] – > VIF; IF << VMAIN}</pre>	(21.147)

(21.147) captures occurrences of the structure with the if element preceding the main verb.

<pre>%14 – IF – after pattern{IF[upos = SCONJ, lemma = “X”]; VIF[upos = VERB AUX]; VMAIN[upos = VERB AUX]; VIF – [mark] – > IF; VMAIN – [advcl] – > VIF; IF >> VMAIN}</pre>	(21.148)

(21.148) captures occurrences of the structure with the if element after the main verb.

Therefore, in each language, we must manually replace the *X* of the lemma with the word (or words) that refer to the *IF* of that language. The following question is obvious: How can such a

lemma be obtained homogeneous, reliable, and efficient? Language To this end, we propose a new methodology, already practiced in the study on colors (since it is lexical research) Brosa-Rodríguez and Jiménez-López (2024) and also recently validated in a similar study (McCarthy, Wu, Mueller, Watson, & Yarowsky, 2020): the use of PanLex (Kamholz, Pool, & Colowick, 2014).

This tool allows us to have these tags that we need in the lemmas in practically all languages with documentation worldwide. To do so, simply enter the equivalent in English and see the results obtained. To avoid homonymy problems, we perform a simultaneous double translation process, which allows us to see if the tag is shared for different meanings. However, this is not too much of a problem since we are restricting to that lemma being a *SCONJ* and, therefore, it is automatically disambiguated from other parts of speech with the same lemma. Moreover, thanks to the list of tag similarities offered by the tool, we can know if in some languages there is not only one lemma for this *IF* but two or more, as it happens in some languages. If the similarity is higher than 66.66% (following the statistical criterion of double or more), the second label in the ranking is also considered for the analysis, as it can be, in the case of Polish, for example, “jeśli” and “jeżeli”.

As usual, we maintain the thresholds already established and statistically supported. If a language has a proportion of 66.66% or more of *IF* before *VMAIN*, we can declare that its normal order is to put the condition before the conclusion. If a language has a proportion of 33.33% or more of *IF* before *VMAIN*, we may declare that its normal order is to put the conclusion before the condition. If the proportion ranges between 33.33% and 66.66%, we should declare that the language has no predominant order.

With all the above, it is possible to quantify all the utterances of real texts of the different languages of the world that contain an “If clause” in Greenberg’s terminology and, in addition, to classify them in the initial position or final position concerning the other element. The formalization proposal allows us to overcome the problems of the complete utterance in its context containing other previous or subsequent elements since, with our pattern, we can perfectly quantify whether the *IF* is previous or after the “conclusion”. However, we want to go further and try to discover a new double Universal or two Universals. We wonder what happens with adverbial clauses in general, without being of the *IF* type. Do they show the same tendency as IF-clauses? A different one? In the second Universal, we want to investigate the existence of some kind of relationship between the distribution ratio of conditional clauses and adverbial clauses in general, whether there is some kind of internal constraint or pressure in the languages. For this purpose, we propose the patterns (21.149) and (21.150):

$$\begin{array}{l}
\%14 - ADVcl - before \\
pattern\{SCONJ[upos = SCONJ]; Vadv[upos = \\
VERB|AUX]; VMAIN[upos = VERB|AUX]; \\
Vadv - [mark] - > SCONJ; VMAIN - [advcl] - > Vadv; SCONJ << \\
VMAIN\}
\end{array} \tag{21.149}$$

(21.149) captures occurrences of adverbial clauses before the main verb.

$$\begin{array}{l}
\%14 - ADVcl - after \\
pattern\{SCONJ[upos = SCONJ]; Vadv[upos = \\
VERB|AUX]; VMAIN[upos = VERB|AUX]; \\
Vadv - [mark] - > SCONJ; VMAIN - [advcl] - > Vadv; SCONJ >> \\
VMAIN\}
\end{array} \tag{21.150}$$

(21.150) captures occurrences of adverbial clauses after the main verb.

The fundamental difference in the patterns we have shown is that we have eliminated *lemma* = “X” concerning the previous patterns since we simply want to eliminate the restriction that the *SCONJ* be of type *IF* and be any. However, to be consistent and more descriptive with the formalization, we renamed the labels we assigned to be more representative. Thus, in these patterns, the *IF* with a value of *SCONJ* we now call directly *SCONJ*; what we previously named *VIF* is now *Vadv* (the verb of the adverbial clause), and we keep the nomenclature to *VMAIN*, which remains unchanged. The dependencies do not change either. We also provide the two possible orders: “*SCONJ* << *VMAIN*” and “*SCONJ* >> *VMAIN*”. As mentioned above, these results can be processed individually to obtain the information discussed in the first Universal, or they can be compared proportionally with the results of the original Universal to see if there is any trend among all the patterns created.

In addition, if we wanted to, we could further expand the possible patterns we can propose to understand better the functioning of structures closely related to conditional clauses or adverbial clauses in general. For example, we can see whether the behavior of adverbial clauses is related to the behavior of adverbs (non-clausal). The prototypical and most closely related relation is “*V* – *mod* – *ADV*”. We can also see if it extends to more parts of speech, such as “*ADJ* – *mod* – *ADV*” or even “*ADV* – *mod* – *ADV*”, where we assume that the behavior is the same or similar. Finally, we can also check the behavior of adverbs in general, regardless of the part of speech they modify, so that we can extend our claims from the most prototypical relations to any possible relation.

The patterns we propose for such an investigation would be (21.151) - (21.156) :

$$\begin{array}{l} \%14 - ADV - beforeV \\ pattern\{V[upos = VERB]; V - [mod]- > ADV; ADV[upos = \\ ADV]; V >> ADV\} \end{array} \quad (21.151)$$

(21.151) captures occurrences of adverbs preceding the verbs they modify.

$$\begin{array}{l} \%14 - ADV - afterV \\ pattern\{V[upos = VERB]; V - [mod]- > ADV; ADV[upos = \\ ADV]; V << ADV\} \end{array} \quad (21.152)$$

(21.152) captures occurrences of adverbs after to the verbs they modify.

$$\begin{array}{l} \%14 - ADV - beforeADJ \\ pattern\{ADJ[upos = ADJ]; ADJ - [mod]- > ADV; ADV[upos = \\ ADV]; ADJ >> ADV\} \end{array} \quad (21.153)$$

(21.153) captures occurrences of adverbs preceding the adjectives they modify.

$$\begin{array}{l} \%14 - ADV - afterADJ \\ pattern\{ADJ[upos = ADJ]; ADJ - [mod]- > ADV; ADV[upos = \\ ADV]; ADJ << ADV\} \end{array} \quad (21.154)$$

(21.154) captures occurrences of adverbs after the adjectives they modify.

$$\begin{array}{l} \%14 - ADV - beforeANY \\ pattern\{ANY[upos = ADJ|ADV|INTJ|NOUN|PROPN| \\ VERB|ADP|AUX|CCONJ|DET|NUM|PART| \\ PRON|SCONJ|SYM]; ANY - [mod]- > ADV; ADV[upos = \\ ADV]; ANY >> ADV\} \end{array} \quad (21.155)$$

(21.155) captures occurrences of adverbs preceding any part of speech they modify.

$$\begin{array}{l}
\%14 - ADV - after ANY \\
pattern\{ANY[upos = ADJ|ADV|INTJ|NOUN|PROPN| \\
VERB|ADP|AUX|CCONJ|DET|NUM|PART| \\
PRON|SCONJ|SYM]; ANY - [mod]- > ADV; ADV[upos = \\
ADV]; ANY << ADV\}
\end{array} \tag{21.156}$$

(21.156) captures occurrences of adverbs after any part of speech they modify.

All these patterns are made in a very similar way. In (21.151), we look for verbs (called *V*) that are the head of an adverb (called *ADV*) that modifies them. We ask that the order of these two elements be “*V* >> *ADV*”, that is, that the adverb appears before the verb. In (21.152), we repeat the same pattern but require that the order of the two elements be the opposite (“*V* << *ADV*”). In the case of (21.153) and (21.154) we repeat the above patterns respectively, but changing *V* to *ADJ* and “*upos = VERB*” to “*upos = ADJ*”. In the last case, (21.155) and (21.156), we repeat the same pattern scheme and change *V* to *ANY* and “*upos = VERB*” to “*upos = ADJ|ADV|INTJ|NOUN|PROPN|VERB|ADP|AUX|CCONJ|DET|NUM|PART|PRON|SCONJ|SYM*”.

21.14.2 Trichotomous Expression Based on Grew-Match Data and Typometrics

This Universal has been specially treated in a publication that is currently being developed, in which both claims of how we have carried out the tests of the rest of the Universals are collected, and new alternative testing methods are explored as an avenue that can be developed in future research.

First of all, we must point out that when we obtained the results, we found that the varieties of Norwegian behaved in an anomalous way, being the only exception in some specific metrics of this Universal (having more occurrences of IF-after than before, while other languages of its genus, type, and area did not show it). Therefore, we reviewed the cases obtained in Norwegian with the help of a Norwegian linguist (Eskil Blaafat Mundal). We concluded that the anomalous results were due to an error in the annotation of Norwegian treebanks.

Two words trigger a condition in the Bokmaal variety: *om* and *hvis*. In the Nyorsk variety, one word triggers a condition: *om*. The problem only occurs with *om*, not with *hvis*. We have detected that these data, contrary to the rest of the languages, are due to 2 different united reasons.

The first error is due to a lack of information that we would like to take advantage of to ask (and we do not know if there is also an annotation error, and it would correspond to another part of speech, which would make the process even easier). When *om* appears in a locution marked by another word, such as *selv*, for example, it ceases to have the meaning of ‘conditional’ and becomes ‘concession,’ as could be “although” in English. We believe that if the information of the type of *SCONJ* it is were added, we could obtain much more attractive and accurate results, not only in this case but in general. As we have discussed, there may also be a need for *selv om* to be labeled in a more representative way and not as *SCONJ*.

Second, we can speak of an annotation error in the treebanks. The form *om* also corresponds to an *ADP* (*PART* if one wants to be less precise and more general), which in English would be equivalent to *about*. Therefore, in this case, the part of speech is fundamental to differentiate these cases; we are facing a disambiguate homonymy thanks to the part of speech, which cannot be performed in this case. We have verified that the treebanks have been automatically converted to the UD scheme and that, therefore, there is a high probability that this indiscriminate *SCONJ* assignment is due to this automatic process, which, with a human review, can be easily solved.

In short, these two cases mentioned above generate a bias that produces misleading results because we look for unwanted structures when we perform the query. It is also necessary to mention that there is always the possibility that similar cases have occurred in other languages since we do not know the hundreds of languages analyzed, and, as can be seen, there are errors in the annotation. However, once the results obtained in the rest of the languages are checked, they are homologous and do not present any apparent problem. In other words, these problems do not exist or, if they do, are so marginal that they do not influence the study.

However, the researcher who is faced with this phenomenon must decide how to correct this error. There are two main options. One option would be to compensate for the error in treebank annotation by negatively filtering out the cases in which *SCONJ* – *mark* – *ADV* (*selv om*) structures appear. The other option is not to search for results based on *om* and to search only with *hvis*, a word without annotation error.

We have chosen the second option for two main reasons. First, although we can separate the cases of *om* equivalent to ‘concessive’ from the cases of *om* as ‘conditional,’ we cannot technically separate the *om* ‘conditional’ from the *om* equivalent to *about*, which is another part of speech but mislabeled.

Secondly, we believe it is unreasonable to modify a formalization to adapt a specific language. This is especially relevant, on the one hand, because we are trying to make a formalization as Universal as possible, applied to all world languages. On the other hand, we are not facing a problem that is intrinsic to the language, but a problem of the annotation, which we believe that a formalization should not try to fix, but that it is much more respectful to wait for the treebank managers to fix the error of assigning part of speech to *om* and, that this word can be included in a more natural way to the model in the future, without searching for *ad hoc* artifices.

Therefore, we discard any occurrence with *om* and simply evaluate the cases with *hvis*, a variety that does not present homonymy and is immune to annotation errors in the POS assignment.

Once this methodological problem has been clarified and we have commented on the decision taken, we can now show the results obtained. The Universal in the terms formulated by Greenberg applies correctly to 59% of the languages analyzed. Since there is a certain link between conditionals and adverbial sentences in UD, we wanted to know the behavior of these and see what the result of this Universal would be without the restriction of dealing with conditional sentences: 48%. Since this result does not seem very satisfactory, we have observed invariants depending on the linguistic type. This has allowed us to draw different hypotheses that translate into Universals that can be tested.

First, we have noticed that languages whose basic word order is SVO tend to place adverbial sentences after the main sentence as the dominant order. We have found this to be true in 91% of the cases. On the other hand, we have proposed the Universal opposite for the other type. That is: “If a language has SOV as its basic order, it will place adverbial sentences before the main sentence”. In this case, the fulfillment is 92%.

Returning to the conditional structures, which do not seem well captured by Greenberg, we observe a tendency to appear more times before the conclusion than the other way around. However, the results obtained (59%) correspond to an inadequate, strict formulation. If we reformulate the Universal in a more Typometric and quantitative way (In conditional statements, the conditional clause precedes the conclusion more than vice versa in all languages), we obtain a compliance of 100%, as shown in the figure 21.29. However, within such a check, we observe that the behavior of the languages is unequal (within the preference of all languages to precede more conditionals than conclusions). The head-final (OV) languages tend to be much more extreme with this conditional antecedent direction. At the same time, the head-initial (VO) shows greater flexibility within the preference for placing IF clauses before usually, as seen in figure 21.30.

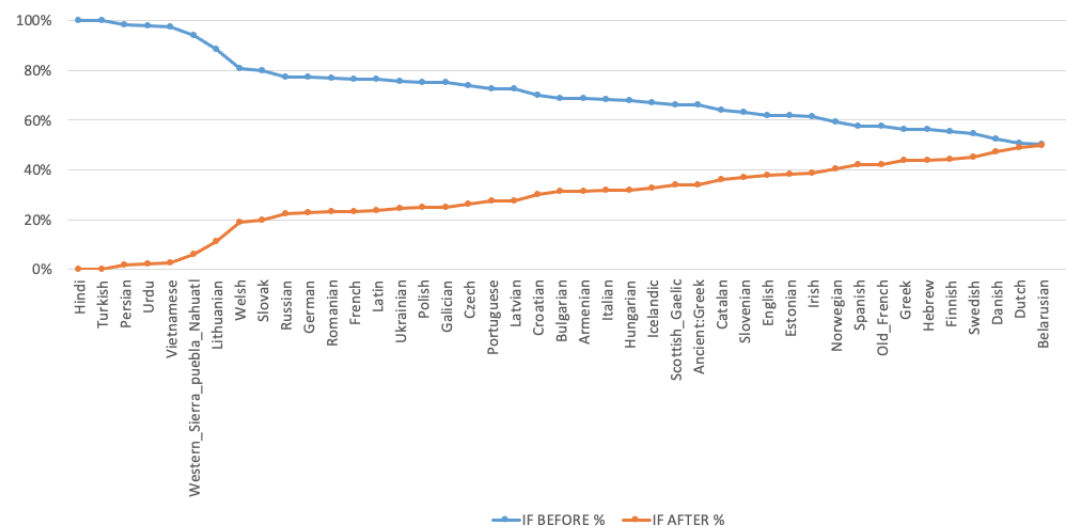


Figure 21.29: Percentage of IF-Clauses before Conclusions

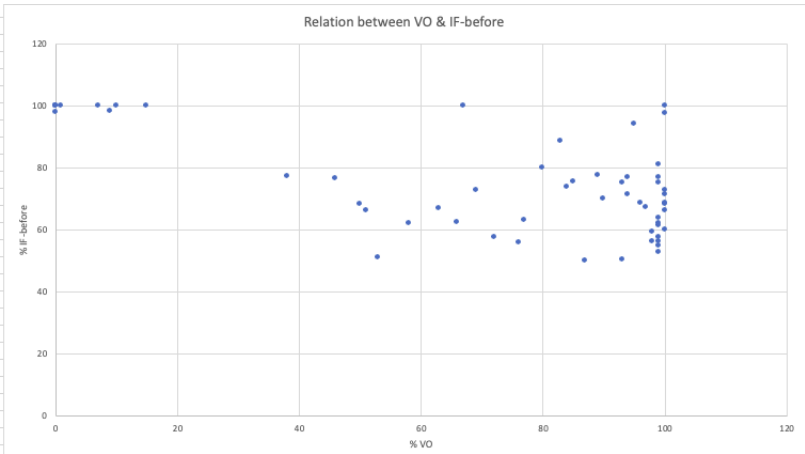


Figure 21.30: Dispersion of Percentage of VO Structures and IF-before Structures

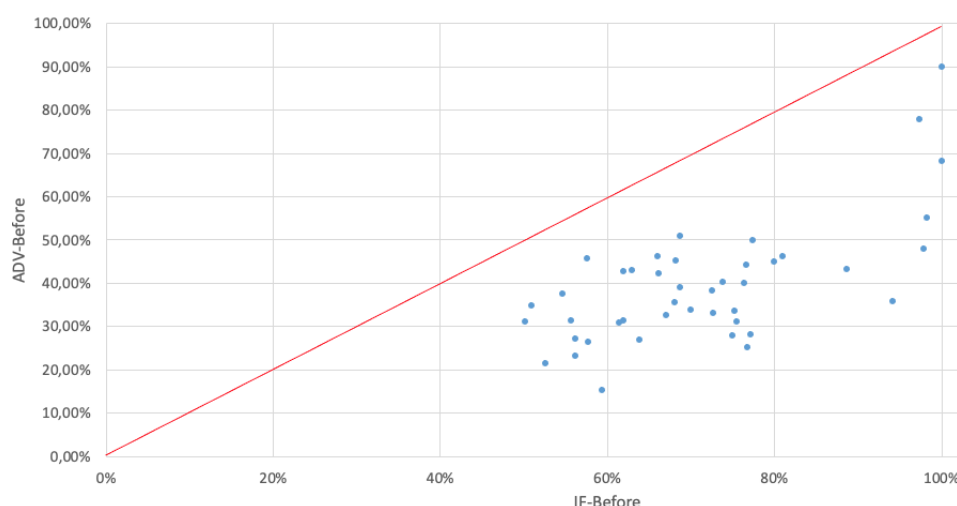


Figure 21.31: Dispersion of Percentage of IF-Clauses before and ADV-Clauses before

Finally, we also wanted to relate the information obtained with the conditional structures and compare it with the information related to adverbial sentences. Again, using a quantitative approach, we can obtain a graph (figure 21.31) where it is pretty clear that the fulfillment of the Universal “In all languages, there are more conditional clauses preceding conclusions than adverbial clauses preceding main clauses” is 100%.

All these data that we have presented can be summarized in the trichotomous expression in figure 21.32:

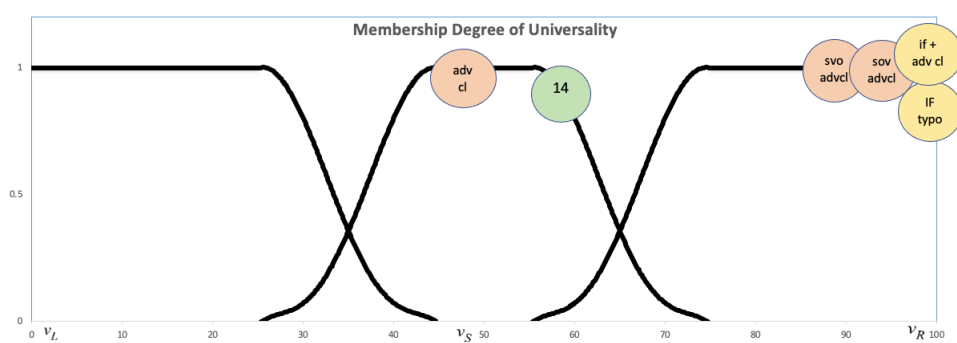


Figure 21.32: Trichotomous Expression for Universal 14

In this case, we can see how the original Universal formulated by Greenberg and the proposal to revise the behavior of conditional sentences are Mid-Universals, indicating the low typological value of such claims in how they are formulated. The formulations of the direction of adverbial clauses conditioned by a given linguistic type (SVO and SOV) are prototypical and High-Universal. Therefore, we recommend the study of adverbial clauses using such a formulation. Greenberg’s

proposed change of the Universal to a less strict, categorical, and more open and quantitative formulation is a prototypical High-Universal. So is the proposed Universal that states that there is a constraint on languages where no matter what proportion of the direction of adverbial sentences this has, conditional sentences will always show a higher occurrence preceding. We can summarise our findings as in table 21.16.

Summary of characteristics of the Universal 14 analysis:
1. All languages show a higher percentage of occurrences of the conditional element before the main verb than after it. It is a High-Universal.
2. OV languages have an extreme placement of the conditional element before the main verb.
3. All languages, regardless of their type or nature, exhibit a higher percentage of preceding conditionals than preceding adverbial sentences. It is a High-Universal.
4. The UD annotation scheme penalizes the labeling of polysynthetic languages and hence their under-representation and the difficulty of capturing conditionals in them.
5. Greenberg’s original Universal has 59% compliance.
6. The same Universal applied to adverbial sentences has a fulfillment of 48%.
7. If we propose the Universal that SVO languages tend to place adverbial sentences after the head in a dominant way, we have a fulfillment of 91%.
8. If we propose the Universal that SOV languages tend to place adverbial preceding sentences dominantly, we have a 92% fulfillment.
9. Greenberg’s original Universal is a Mid-Universal. There is no such dominance, although there is a tendency.
10. The Universal of the tendency to place adverbial clauses preceding is a Mid-Universal.
11. The Universal linking the SVO type with the adverbial clauses placed after the head is a High-Universal.
12. The Universal linking the SOV type with the adverbial clauses preceding is a High-Universal.

Table 21.16: Summary of Characteristics of the Universal 14 Analysis

21.15 Universal 15

UNIVERSAL 15: *In expressions of volition and purpose, a subordinate verbal form always follows the main verb in the normal order, except in languages where the nominal object always precedes the verb.*

21.15.1 Current Limitation: Data no Available

In the case of Universal 15, we find a condition in the implication impossible to know in the utterances available in the UD treebanks. We must know whether such an utterance expresses “volition” or “purpose”. Leaving aside possible pragmatic implications that may affect such an utterance, we do not know any semantic information if we use UD. This does not allow us to know these expressions in the different languages of the world systematically. On the other hand, morphosyntactic elements can express this type of expression. However, these are not Universal, and each language may use different strategies.

We may believe we have similar information since the UD annotation guide provides a mood feature that is “desiderative”, for example. However, this only applies to languages that make a clear difference between two forms to distinguish desiderative from non-desiderative using the verb mood, such as Japanese or Turkish. Thus, we only potentially find the distinction in those languages and will not find any *Mood = Des* in Romance languages, for example, since they use other mechanisms. Even if we perform this search in languages such as Japanese and Turkish, only the latter contains occurrences of this element. In short, it is not technically possible to work with this information.

21.16 Universal 16

UNIVERSAL 16: *In languages with dominant order VSO, an inflected auxiliary always precedes the main verb. An inflected auxiliary always follows the main verb in languages with dominant order SOV.*

21.16.1 How to Check It in Typometrics

Technical Decisions: 1. Firstly, the Universal is analyzed with Typometrics (Gerdes et al., 2021). 2. We select the SUD annotation scheme because we can get results more comparable to the classical typological claims. 3. We select Graph 2D. 4. We set a filter of utterances to avoid misleading data. 5. We have to generate two different graphs since this Universal is twofold; one applies to one linguistic type and the other to another. For the first one, we should select <i>direction – cfc</i> in both x-Axis and y-Axis, meaning that we are interested in the word order feature but applied only in a specific Part of Speech. 6. in the x-Axis, we should select <i>VERB – subj – NOUN</i> (nominal subjects, since Greenberg requires basic word order). In the y-Axis, we select <i>AUX – comp : aux – VERB</i> (the way the relation between a verb and auxiliary is represented in SUD format). 7. For the second plot, in x-Axis we place <i>VERB – comp : obj – NOUN</i> and in y-Axis we place <i>AUX – comp : aux – VERB</i>.

Table 21.17: Technical Decisions for Checking Universal 16 in Typometrics

We must check by WALS or Grew-Match if the OV languages we detected correspond to SOV, not OVS or OSV, which is rare. Again, we apply this Universal by using the data after the checking done in U4 (21.3), that is, Akkadian, Amharic, Bhojpuri, Bambara, Buryat, Basque, Persian, Hindi, Western Armenian, Kazakh, Kurmanji, Korean, Marathi, Turkish German, Tamil, Telugu, Turkish, Uyghur, and Urdu.

A problem we find with Typometrics is our morphological limitations since the tool does not allow greater customization, as we have already mentioned when we presented Typometrics in the previous chapter. This means that we cannot filter the auxiliaries by those that have inflection from those that do not. Therefore, in Typometrics we must give up making such a distinction and try to see if the results without such detail still form a Universal, a fact that would make it possible to slightly modify the Greenberg Universal. On the other hand, such assessments can be made by Grew-Match (a more conservative or restricted approach), as we will show below with different tests.

21.16.2 Formalization Proposed for Grew-Match

In Grew-Match, we must select all the languages that have a dominant order VSO for the first part of the Universal and SOV for the second part of the Universal. To do this, we must, as we have already mentioned, look for the basic word order of all the languages, make a ranking of occurrences, and see if the frequency of the first order is double that of the second to determine that this language has a dominant order (and it is not free order). In conclusion, we must determine the languages to which we will apply the first part and the second part of the Universal, so we must propose the patterns (21.157) - (21.162):

$$\begin{array}{l}
 \%16 - SVO \\
 pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = \\
 PROPEN|NOUN]; V - [comp : obj] - > O; O[upos = \\
 PROPEN|NOUN]; V >> S; S << O; V << O\} \\
 without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
 \end{array} \tag{21.157}$$

(21.157) captures the SVO word order. We ask the subject to be before the verb and the object to be after the verb.

$$\begin{array}{l}
 \%16 - SOV \\
 pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = \\
 PROPEN|NOUN]; V - [comp : obj] - > O; O[upos = \\
 PROPEN|NOUN]; V >> S; S << O; V >> O\} \\
 without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
 \end{array} \tag{21.158}$$

(21.158) captures the SOV word order. We ask the subject to be before the verb and the object before the verb but not before the subject.

$$\begin{array}{l}
 \%16 - VSO \\
 pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = \\
 PROPEN|NOUN]; V - [comp : obj] - > O; O[upos = \\
 PROPEN|NOUN]; V << S; S << O\} \\
 without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
 \end{array} \tag{21.159}$$

In the VSO case, in (21.159), we want a verb before the subject and before the object, but the subject before the object.

$$\begin{array}{l}
 \%16 - VOS \\
 pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = \\
 PROPEN|NOUN]; V - [comp : obj] - > O; O[upos = \\
 PROPEN|NOUN]; V << S; S >> O; V << O\} \\
 without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
 \end{array} \tag{21.160}$$

In VOS word order, in (21.160), we formalize the restriction of verb being before the subject and before the object, but the object appearing before the subject.

$$\begin{array}{l}
 \%16 - OSV \\
 pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = \\
 PROPEN|NOUN]; V - [comp : obj] - > O; O[upos = \\
 PROPEN|NOUN]; V >> S; S >> O; V >> O\} \\
 without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
 \end{array} \tag{21.161}$$

In the OSV order, in (21.161), we ask for objects appearing before subjects and verbs, but subjects before verbs.

$$\begin{array}{l}
 \%16 - OVS \\
 pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = \\
 PROPEN|NOUN]; V - [comp : obj] - > O; O[upos = \\
 PROPEN|NOUN]; V << S; S >> O; V >> O\} \\
 without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
 \end{array} \tag{21.162}$$

Moreover, in OVS order, in (21.162), we ask for objects appearing before subjects and verbs, but verbs before subjects.

For the conclusion of the implicative Universal, we must formalize the relation of a verb to an auxiliary only when the latter is inflected. Therefore, we decided to name an element *AUX* corresponding to a *group* of *AUX* and the feature of *VerbForm* = *Fin*. This is because under this nomenclature are grouped all the auxiliaries (or verbs) that in traditional terminology would be considered inflected and have mood information. We call the second element of the pattern *V*, corresponding to the *upos* of *VERB*. Then, we create the dependency relation, where we place the *AUX* as head and the *V* as dependent (since in SUD they have decided to assign the *AUX* in this type of situation the role of head) and where we look for a dependency relation of *comp* : *aux*, which is the typical relation existing between an auxiliary and any element that complements it. Finally, we must indicate the two orders in which they can appear. In one pattern, we must include the order *AUX* << *V*, where the *AUX* would come before the *V*. In the other pattern, we must include the order *AUX* >> *V*, where the *V* would come before the *AUX*. This development of formalization we have explained can be seen in the patterns (21.163) and (21.164).

$$\begin{array}{l} \%16 - AUX - after - fin \\ pattern\{AUX[upos = AUX, VerbForm = Fin]; AUX - [comp : \\ aux] - > V; V[upos = VERB]; AUX >> V\} \end{array} \quad (21.163)$$

(21.163) captures finite auxiliaries after the depending verb.

$$\begin{array}{l} \%16 - AUX - before - fin \\ pattern\{AUX[upos = AUX, VerbForm = Fin]; AUX - [comp : \\ aux] - > V; V[upos = VERB]; AUX << V\} \end{array} \quad (21.164)$$

(21.164) captures finite auxiliaries before the depending verb.

However, we do not want to be simply verifying what Greenberg proposed. It is striking that he has decided to restrict this Universal to inflected auxiliaries. Therefore, we want to change different parameters around this feature to represent the behavior of *AUX* with other features and to be able to nuance Greenberg's Universal or to create new Universals.

The first case we want to verify is what happens when we do not filter any restriction regarding the *AUX* features, that is, if we take into account, indiscriminately, *AUX* with and without inflection. So, we remove from the above two patterns the feature of *VerbForm* = *Fin*. That is, what we present in the patterns (21.165) and (21.166).

$$\begin{array}{l} \%16 - AUX - after \\ pattern\{AUX[upos = AUX]; AUX - [comp : aux] - > V; V[upos = \\ VERB]; AUX >> V\} \end{array} \quad (21.165)$$

(21.165) captures auxiliaries after the depending verb.

$$\begin{array}{l} \%16 - AUX - before \\ pattern\{AUX[upos = AUX]; AUX - [comp : aux] - > V; V[upos = \\ VERB]; AUX << V\} \end{array} \quad (21.166)$$

(21.166) captures auxiliaries after the depending verb.

Another option is to check the behavior of *AUX* that are not infinitives. We do not know whether the isolation of this single variety may be what Greenberg was concerned about, and this distinction may yield different or more illuminating results. So, to the *AUX*, we add the feature of *VerbForm* $\neq$ *Inf*, which what it does is to take into account any *AUX* that does not have the value of *Inf*. This can be seen in the patterns (21.167) and (21.168).

$$\begin{array}{l} \%16 - AUX - after - nINF \\ pattern\{AUX[upos = AUX, VerbForm \neq Inf]; AUX - [comp : \\ aux] - > V; V[upos = VERB]; AUX >> V\} \end{array} \quad (21.167)$$

(21.167) captures non-infinitive auxiliaries after the depending verb.

$$\begin{array}{l} \%16 - AUX - before - nINF \\ pattern\{AUX[upos = AUX, VerbForm \neq Inf]; AUX - [comp : \\ aux] - > V; V[upos = VERB]; AUX << V\} \end{array} \quad (21.168)$$

(21.168) captures non-infinitive auxiliaries before the depending verb.

Moreover, we can check what happens when we extend this restriction to the three typical types in classical terminology: infinitives, gerunds, and participles. Predictably, the behavior excluding this group will be similar to excluding only the infinitive. Therefore, we can check the weight of gerunds and participles. We only have to add to *AUX* the feature *VerbForm* $\langle \rangle$ *Inf|Part|Ger*, which will exclude these three options. The patterns would be (21.169) and (21.170).

$$\begin{array}{l} \%16 - AUX - after - nINFpartger \\ pattern\{AUX[upos = AUX, VerbForm \langle \rangle Inf|Part|Ger]; AUX - \\ [comp : aux] - > V; V[upos = VERB]; AUX >> V\} \end{array} \quad (21.169)$$

(21.169) captures non-infinitive/participles/gerunds auxiliaries after the depending verb.

$$\begin{array}{l} \%16 - AUX - before - nINFpartger \\ pattern\{AUX[upos = AUX, VerbForm \langle \rangle Inf|Part|Ger]; AUX - \\ [comp : aux] - > V; V[upos = VERB]; AUX << V\} \end{array} \quad (21.170)$$

(21.170) captures non-infinitive/participles/gerunds auxiliaries before the depending verb.

Finally, we propose a last option corresponding to the negative version of the original pattern, that is, to exclude all *Fin*, which was the only thing we were looking for before. That is, we propose to add to *AUX* the feature *VerbForm* $\langle \rangle$ *Fin*. The proposed patterns would be (21.171) and (21.172).

$$\begin{array}{l} \%16 - AUX - after - nFINpattern\{AUX[upos = AUX, VerbForm \langle \rangle \\ Fin]; AUX - [comp : aux] - > V; V[upos = VERB]; AUX >> V\} \end{array} \quad (21.171)$$

(21.171) captures non-finite auxiliaries after the depending verb.

$$\begin{array}{l} \%16 - AUX - before - nFINpattern\{AUX[upos = \\ AUX, VerbForm \langle \rangle Fin]; AUX - [comp : aux] - > V; V[upos = \\ VERB]; AUX << V\} \end{array} \quad (21.172)$$

(21.172) captures non-finite auxiliaries before the depending verb.

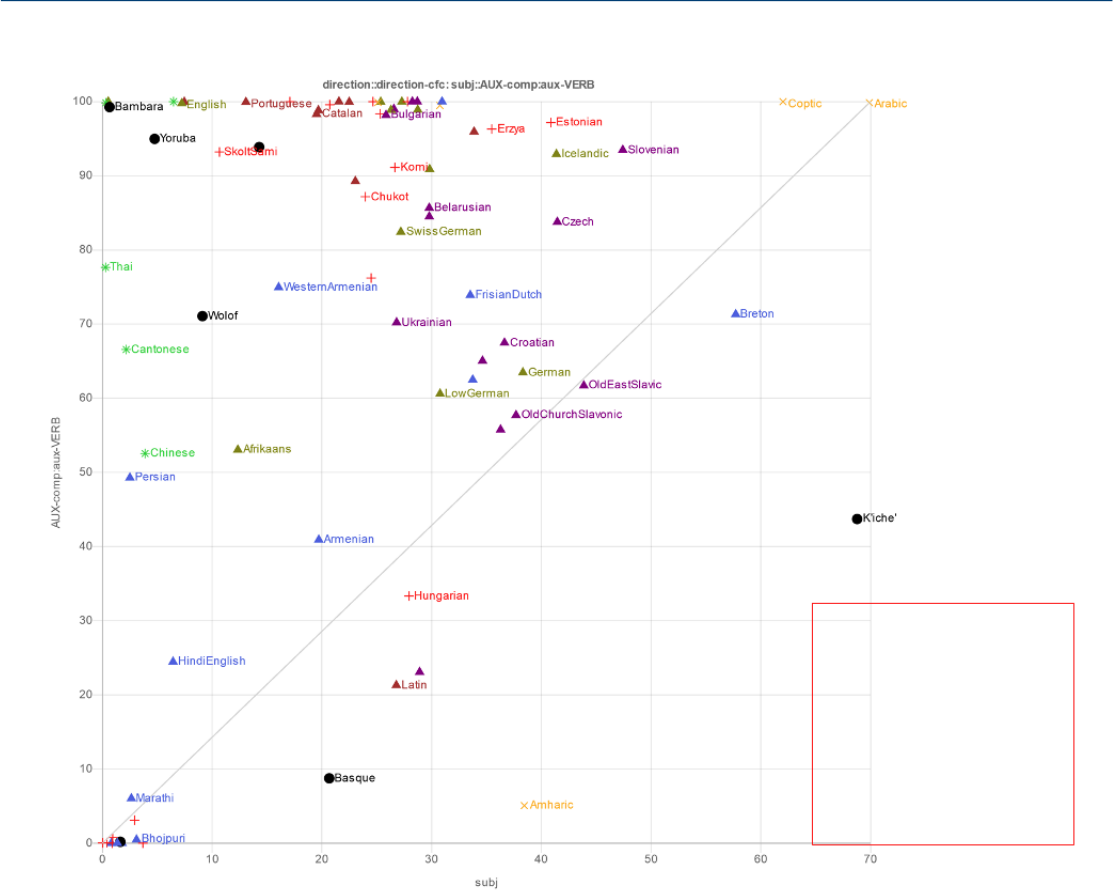


Figure 21.33: Subjects and Auxiliaries Direction Checking

21.16.3 Graphs from Typometrics

This Universal is composed of two different Universals. The first one concerns languages with linguistic type VSO. The result of the analysis in Typometrics can be seen in the figures 21.33 and 21.34.

As can be seen, the red square is empty, so we can determine that this part of the Universal proposed by Greenberg is true. However, we must underline why there is no language in such a square: the graph stops at 70% on the horizontal axis.

This means there are no VSO languages (beyond Arabic, for example, which appears on the boundary) that have occurrences of auxiliary forms dependent (or heads in the case of SUD) on verbs. Thus, rather than an absence at the location designated by Greenberg, there appears to be a general absence of this feature in languages of this linguistic type. However, the low sample of languages of this linguistic type and the shallow depth of knowledge we have of languages of this type means that we simply limit ourselves to pointing out this possibility, something that may be relevant to contrast with data at another level (grammars, for example) and with a larger sample.

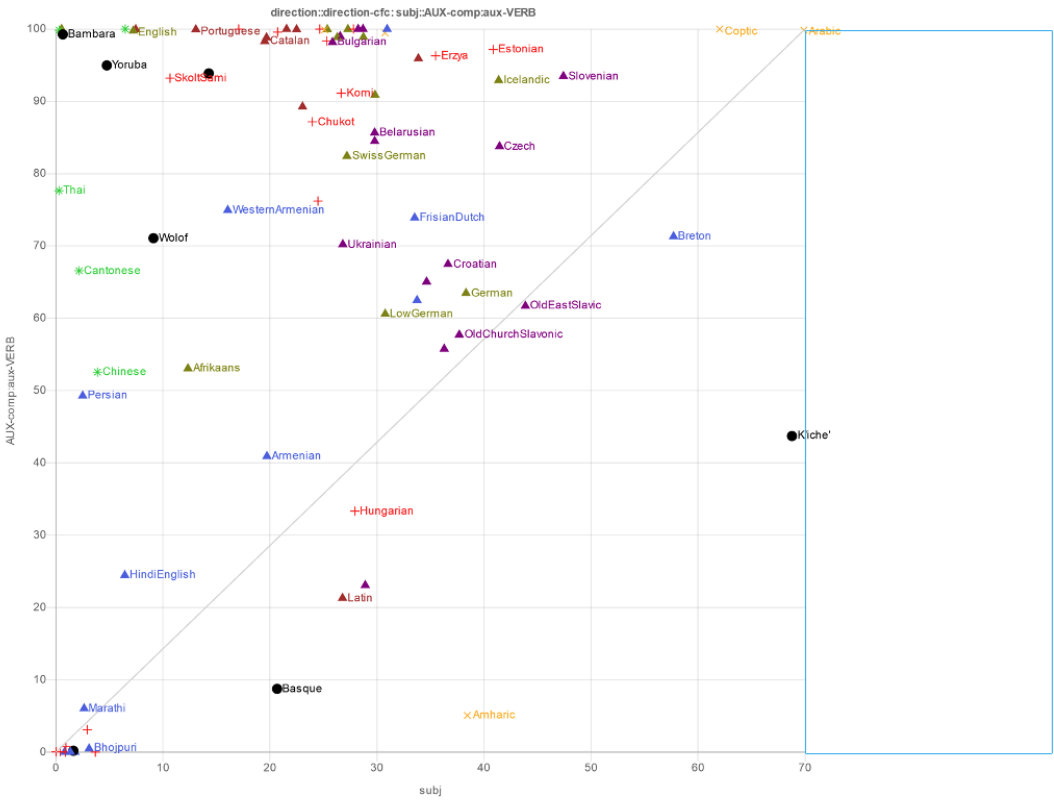


Figure 21.34: Subjects and Auxiliaries Direction Proposal

No significant forms or correlations of language groups can be seen in such a graph, so in figure 21.34, we limit ourselves to highlighting with a blue rectangle the area of the graph that is not represented because it does not show any occurrence.

The second part of the Universal concerns languages of type SOV. In this case, represented in the figure 21.35.

These languages (on the left side of the graph) should place the verb-dependent auxiliaries (or heads in SUD) after the verb. In other words, these languages will have a value of 0 or close to 0 for verbs after the auxiliary and, therefore, will be placed in the lower left corner. As seen in the red box that we propose to mark the space that should appear empty according to Greenberg, we observe different languages appearing there. These languages have different characteristics and appear at different points in the box. However, it is necessary to count more precisely the weight they have in proportion to the SOV languages in general.

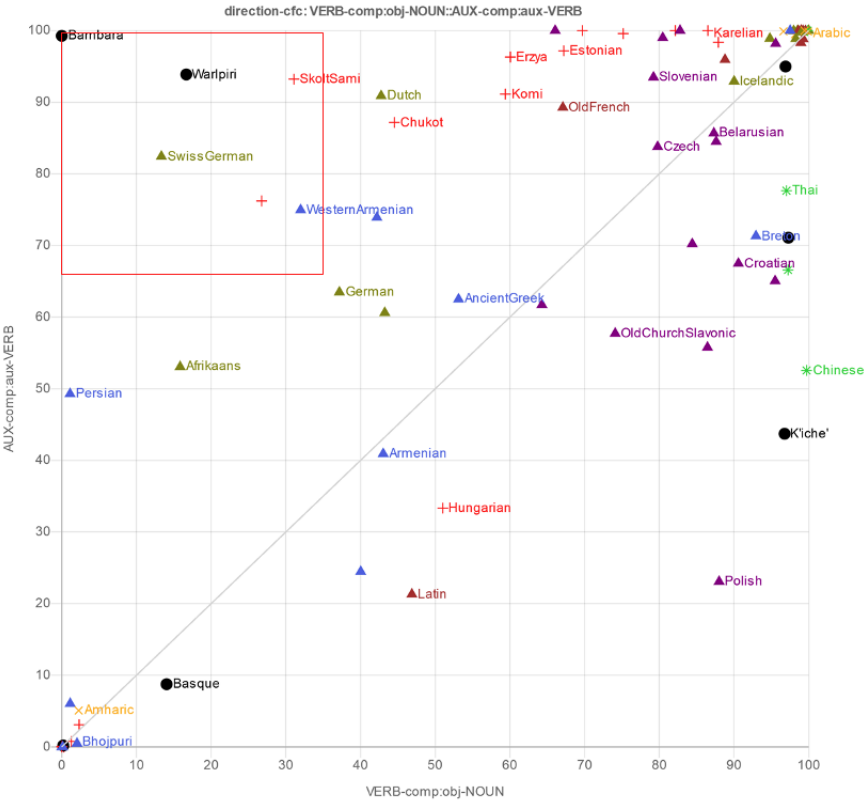


Figure 21.35: Objects and Auxiliaries Direction Checking

21.16.4 Trichotomous Expression Based on Grew-Match Data

The findings are:

- The overall Universal (taking into account inflected auxiliaries) has a compliance of 80%. This 80% is the result of considering the results obtained for the part of the Universal applicable to SOV languages since we did not carry out such a check due to the lack of occurrences in different VSO languages,
- We also propose to check a new Universal that would be the relation between the linguistic type SVO and the inflected auxiliaries preceding the verbs, with a fulfillment of 90%.
- On the other hand, we have also been interested in pondering the weight that the restriction added by Greenberg that such auxiliaries must be inflected has in such a Universal. For this purpose, we carried out the check without such a restriction in order to oppose the result proposed by Greenberg with this one. In such a case, we obtain the same result every time in the checks, with an 80% in the case of SOV, for example. This means that the limitation proposed by Greenberg is unnecessary and could be eliminated for greater extension and representativeness.

Once we have such data, we can propose a trichotomous expression on this Universal, as shown in figure 21.36.

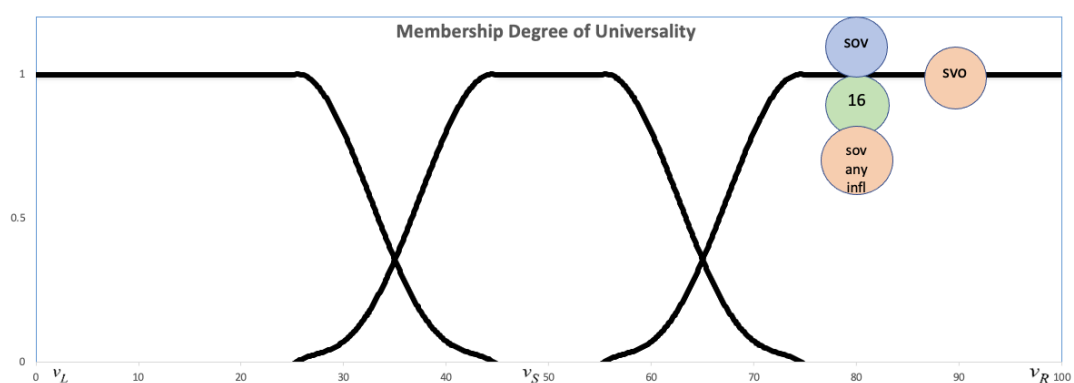


Figure 21.36: Trichotomous Expression for Universal 16

In this case, we have placed the checks of the proposals: original (green), SOV type (blue), SVO type (orange right), and SOV type without inflection restriction (orange left). As can be seen, in all cases, the checks are located in the High-Universal group (and at the maximum value

of representativeness of this group). Therefore, it seems that Greenberg’s proposal and our two proposals are interesting when it comes to understanding the functioning of the world’s languages. However, as we have already commented, we believe that the behavior of VSO languages, which has been discarded in this Universal for lack of occurrences, should be tested in greater depth and extension. We can summarise our findings as in table 21.18.

Summary of characteristics of the Universal 16 analysis: 1. Typometrics: There are practically no occurrences of SOV languages with the use of auxiliaries (right-hand side of the first pair of graphs). 2. Typometrics: We find languages such as Warlpiri, Bambara, or Sami, which are exceptions to the SOV part of the Universal (upper left corner of the second pair of graphs). 3. The original Greenberg Universal, in all its terms, has a compliance of 80 4. If the “finite auxiliaries” restriction is removed and they are merely “auxiliaries”, the compliance is the same. 5. The proposal to characterize SVO-type languages as languages that tend to place auxiliaries preceding verbs has a 90 6. Greenberg’s original Universal is a High-Universal. 7. The Universal we propose for SVO languages is a High-Universal. 8. The Universal that removes the restriction that auxiliaries be finite is a High-Universal.

Table 21.18: Summary of Characteristics of the Universal 16 Analysis

21.17 Universal 17

UNIVERSAL 17: *With overwhelmingly more than chance frequency, languages with dominant order VSO have the adjective after the noun.*

21.17.1 How to Check It in Typometrics

Technical Decisions:

1. Firstly, the Universal is analyzed with Typometrics (Gerdes et al., 2021).
2. We select the SUD annotation scheme because we can get results more comparable to the classical typological claims.
3. We select Graph 2D.
4. We set a filter of utterances to avoid misleading data.
5. We select *direction* and the *subj* dependency in the x-Axis.
6. We select *direction – cfc* and *NOUN – mod – ADJ* in the y-Axis. It corresponds to adjectives modifying nouns, which is what Greenberg postulated.

Table 21.19: Technical Decisions for Checking Universal 17 in Typometrics

In this case, we have six languages that could be placed as VS and, therefore, potentially VSO. These languages are Manx, Gaelic, Irish, Welsh, Arabic, and Kiche. However, we decided to discard Maya Kiche. In WALS, there is no data on its basic word order. Once we check it in Grew-Match, we see that its behavior is closer to the non-dominant order (with a tendency to be VSO). It should also be noted that Greenberg, Welsh has analyzed only one of these languages.

It is unnecessary to make any further checks, as this Universal is simple. We simply need to check the behavior of the mentioned languages in relation to the order of their adjective-modifying nouns (from the threshold established in this thesis in previous Universals), which is what will be shown in the results.

However, we will not only look at the distribution in the VSO languages, which is the one indicated by Greenberg, but we will also see if we can identify any other homogeneous behavior in

other linguistic types (although we will also have to develop some of this in Universal 19, which deals with the same dependency). In addition, we should also say that it would be interesting to see if the detected invariants are particular to the *mod* dependency between adjectives and nouns or, indeed, if a similar approach occurs in other dependencies.

For this, we have decided to use a functionality with quite a lot of potential of Typometrics (Gerdes et al., 2021) that, to our knowledge, has not yet been exploited by anyone, given the lack of knowledge: Closest Graph Function. This function compares the distributions of all the graphs that it is possible to generate to see in which of the dependencies there is a smaller variation of the individual position of each language or which graph is closer to ours (it only works with 1D graphs). By closest, we mean counting how far or close a language moves in the scale. For example, suppose a language is 86% *NOUN – mod – ADJ* and 83,5% *ADJ – mod – ADV* (which happens more or less with every language). In that case, both distributions behave the same way since there is no significant difference in the movement of every singular language. This functionality is available and exemplified in Typometrics (Gerdes et al., 2021) for further info.

To use the highlighted functionality, we set up the *NOUN – mod – ADJ* dependency in a 1D graph and check the closest graph sorted from closest to least. After checking the first ones, which are irrelevant because they are obvious, such as the dependency on *PROPN – mod – ADJ*, we find quite a few similarities between adjectives as modifiers and adverbs as modifiers. Therefore, we created a 2D graph with the two dependencies, *NOUN – mod – ADJ* and *ADJ – mod – ADV*, on the x-axis and y-axis, respectively, to visually see the correlation.

It is also possible to create a different 2D plot, comparing this newfound dependence with the strongest dependence predictor according to Dryer (2013), the OV-VO opposition. According to Dryer, in this particular case, the adverb as a modifier of an adjective is not a correlation pair of the OV-VO opposition. For this, we must put in the x-Axis *VERB – comp : obj – NOUN* and in the y-Axis, again, *ADJ – mod – ADV*.

21.17.2 Formalization Proposed for Grew-Match

If we want to analyze Grew-Match, first of all, we must determine the basic word order of the languages to know to which ones the Universal applies and to which ones it does not, by means of the methodology already explained in this work. The patterns that would be used are the ones from Universal 1: (21.173) -(21.178).

$$\begin{array}{l}
\%17 - SVO \\
pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = \\
PROPN|NOUN]; V - [comp : obj] - > O; O[upos = \\
PROPN|NOUN]; V >> S; S << O; V << O\} \\
without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
\end{array} \tag{21.173}$$

In (21.173), we capture the SVO word order we ask the subject to be before the verb and the object to be after the verb.

$$\begin{array}{l}
\%17 - SOV \\
pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = \\
PROPN|NOUN]; V - [comp : obj] - > O; O[upos = \\
PROPN|NOUN]; V >> S; S << O; V >> O\} \\
without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
\end{array} \tag{21.174}$$

In (21.174), we capture the SOV word order. We ask the subject to be before the verb and the object before the verb but not before the subject.

$$\begin{array}{l}
\%17 - VSO \\
pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = \\
PROPN|NOUN]; V - [comp : obj] - > O; O[upos = \\
PROPN|NOUN]; V << S; S << O\} \\
without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
\end{array} \tag{21.175}$$

In the VSO case, in (21.175), we want a verb before the subject and before the object, but the subject before the object.

$$\begin{array}{l}
\%17 - VOS \\
pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = \\
PROPN|NOUN]; V - [comp : obj] - > O; O[upos = \\
PROPN|NOUN]; V << S; S >> O; V << O\} \\
without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
\end{array} \tag{21.176}$$

In VOS word order, in (21.176), we formalize the restriction of verb being before the subject and before the object, but the object appearing before the subject.

$$\begin{array}{l}
\%17 - OSV \\
pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = \\
PROP N|NOUN]; V - [comp : obj] - > O; O[upos = \\
PROP N|NOUN]; V >> S; S >> O; V >> O\} \\
without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
\end{array} \tag{21.177}$$

In the OSV order, in (21.177), we ask for objects appearing before subjects and verbs, but subjects before verbs.

$$\begin{array}{l}
\%17 - OVS \\
pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = \\
PROP N|NOUN]; V - [comp : obj] - > O; O[upos = \\
PROP N|NOUN]; V << S; S >> O; V >> O\} \\
without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
\end{array} \tag{21.178}$$

Moreover, in OVS order, in (21.178), we ask for objects appearing before subjects and verbs, but verbs before subjects.

Subsequently, we must check the order of the adjectives modifying nouns. We decide to name an element *ADJ* by assigning it a value of *upos* = *ADJ* and another element *N* by assigning it a value of *upos* = *NOUN|PROP N*. Then, we create the dependency relation. In this case, we want it to be *N* – [*mod*] – > *ADJ*; that is, we want to have *N* as head and *ADJ* as dependent by modifying it. Then, in one pattern, we ask for the *N* to be in the first position (*N* << *ADJ*) and in another, for it to be in the second position (*N* >> *ADJ*). In short, we want the patterns (21.179) and (21.180).

$$\begin{array}{l}
\%17 - ADJ - after \\
pattern\{ADJ[upos = ADJ]; N - [mod] - > ADJ; N[upos = \\
PROP N|NOUN]; N << ADJ\}
\end{array} \tag{21.179}$$

(21.179) captures adjectives after the noun they modify.

$$\begin{array}{l}
\%17 - ADJ - before \\
pattern\{ADJ[upos = ADJ]; N - [mod] - > ADJ; N[upos = \\
PRON|NOUN]; N >> ADJ\}
\end{array} \tag{21.180}$$

(21.180) captures adjectives before the noun they modify.

Another option is to see what behavior these adjectives have when they modify a pronoun and not a noun to see if Greenberg's basic and neutral proposal separates distinct behaviors or, in fact, we can extend these results. Thus, from the previous patterns, we change the assignment of the label N its content, being now $[upos = PRON]$. However, to be consistent, we change the assigned label N to PN . Thus, the head of this relation will now be PN , and also, the first element of which a given order is sought will be PN . We propose the patterns (21.181) and (21.182).

$$\begin{array}{l}
\%17 - ADJ - PRON - after \\
pattern\{ADJ[upos = ADJ]; PN - [mod] - > ADJ; PN[upos = \\
PRON]; PN << ADJ\}
\end{array} \tag{21.181}$$

(21.181) captures adjectives after the pronoun they modify.

$$\begin{array}{l}
\%17 - ADJ - PRON - before \\
pattern\{ADJ[upos = ADJ]; PN - [mod] - > ADJ; PN[upos = \\
PRON]; PN >> ADJ\}
\end{array} \tag{21.182}$$

(21.182) captures adjectives before the pronoun they modify.

Perhaps some researchers are interested in knowing the behavior of adjectives in predicative and not modifier relations. Although we do not believe that clear correspondence can be established with the rest of the patterns, since here the similarity with the OV/VO opposition seems closer, we want to offer the formalization of these situations, which may be helpful to draw future conclusions in more complex studies. In this case, we must make a total change in the patterns; the only thing that remains is $ADJ[upos = ADJ]$. In this case, we remove any N element and instead create a V element with the value of $[upos = VERB]$. Our dependency relation corresponds with the predicative, i.e., $V - [comp : pred] - > ADJ$. Finally, the two possible orders would be $V >> ADJ$ or $V << ADJ$. The final patterns would have the form of (21.183) and (21.184).

$$\begin{array}{l}
\%17 - ADJ - predicative - before \\
pattern\{V[upos = VERB|AUX]; V - [comp : pred] - > \\
ADJ; ADJ[upos = ADJ]; V >> ADJ\}
\end{array} \tag{21.183}$$

(21.183) captures predicatives before their verb.

$$\begin{array}{l}
\%17 - ADJ - predicative - after pattern\{V[upos = VERB|AUX]; V - \\
[comp : pred] - > ADJ; ADJ[upos = ADJ]; V << ADJ\}
\end{array} \tag{21.184}$$

(21.184) captures predicatives after their verb.

Another possibility, which we have already discussed in other Universals, is to remove any kind of restriction concerning the part of speech and only consider the dependency. Thus, we could see if Greenberg's claims are extensible to any type of element or, on the contrary, only be circumscribed to the most prototypical cases in each case. In this example, the prototypical case would be modifying a name. To make this formalization we must replace the element *N* by an element *ANY*, for example. The value we assign to the *ANY* tag would be $[upos = ADJ|ADV|INTJ|NOUN|PROPN|VERB|ADP|AUX|CCONJ|DET|NUM|PART|PRON|SCONJ|SYM]$. Consequently, the dependency relationship is $ANY - [mod] - > ADJ$, an adjective modifying any element. The order can be $ANY << ADJ$ or $ANY >> ADJ$. The complete formalizations would be (21.185) and (21.186).

$$\begin{array}{l}
\%17 - ADJ - any - after \\
pattern\{ANY[upos = ADJ|ADV|INTJ|NOUN|PROPN| \\
VERB|ADP|AUX|CCONJ|DET|NUM|PART|PRON| \\
SCONJ|SYM]; ANY - [mod] - > ADJ; ADJ[upos = ADJ]; ANY << \\
ADJ\}
\end{array} \tag{21.185}$$

(21.185) captures adjectives after any part of speech they modify.

$$\begin{array}{l}
\%17 - ADJ - any - before \\
pattern\{ANY[upos = ADJ|ADV|INTJ|NOUN|PROPN| \\
VERB|ADP|AUX|CCONJ|DET|NUM|PART|PRON| \\
SCONJ|SYM]; ANY - [mod]- > ADJ; ADJ[upos = ADJ]; ANY >> \\
ADJ\}
\end{array} \tag{21.186}$$

(21.186) captures adjectives before any part of speech they modify.

If we want the patterns in Grew-Match of the extra findings made in Typometrics, we must completely change the proposed formalization. In the case of the first extra graph, in which we compare the relation of *NOUN-mod-ADJ* with *ADJ-mod-ADV*, we simply need to formalize the latter, since the former we have already presented previously in (21.179) and (21.180). The patterns we propose are (21.187) and (21.188).

$$\begin{array}{l}
\%17 - ADV - after \\
pattern\{ADJ[upos = ADJ]; ADJ - [mod]- > ADV; ADV[upos = \\
ADV]; ADJ << ADV\}
\end{array} \tag{21.187}$$

(21.187) captures adverbs after the adjective they modify.

$$\begin{array}{l}
\%17 - ADV - before \\
pattern\{ADJ[upos = ADJ]; ADJ - [mod]- > ADV; ADV[upos = \\
ADV]; ADJ >> ADV\}
\end{array} \tag{21.188}$$

(21.188) captures adverbs before the adjective they modify.

In these cases, we have activated two elements: *ADJ[upos = ADJ]* and *ADV[upos = ADV]*. We place the modified object as head (*ADJ*) and the modifier as dependent (*ADV*). Furthermore, we name the dependency *-[mod]- >*. In the first pattern, we ask that the adjective appears in the first position (*ADJ << ADV*) and in the second that the adverb appears in the first position (*ADJ >> ADV*).

If we want to go further, as we have discussed in part on creating graphs with Typometrics, and check whether these modification relationships are a correlation pair or not with the most important relationship (OV/VO), we simply have to reuse the patterns we are interested in contrasting with

this correlation pairs trigger (modifying adjectives or modifying adverbs) and compare them with their distribution with the patterns (21.189) and (21.190).

$$\begin{array}{l}
 \%17 - OV \\
 pattern\{V[upos = VERB]; V - [subj] - > O; O[upos = \\
 PROP N|NOUN]; V >> O\} \\
 without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
 \end{array} \quad (21.189)$$

(21.189) captures objects before their verb.

$$\begin{array}{l}
 \%17 - VO \\
 pattern\{V[upos = VERB]; V - [subj] - > O; O[upos = \\
 PROP N|NOUN]; V << O\} \\
 without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
 \end{array} \quad (21.190)$$

(21.190) captures objects after their verb.

In these patterns, as can be seen, the only thing we do is to take advantage of a specific formalization already developed in Universal 1 (21.1), where we ask that a noun (common or proper) be the direct object of a verb. In one case, we ask for that noun to appear before the verb and, in another, after.

21.17.3 Graphs from Typometrics

In this Universal, again, we recheck the linguistic type VSO and its relation to the position of the adjective (as nominal modifier), as we show in figures 21.37 and 21.38.

VSO languages, located on the right of the figure 21.37, tend to be located in the highest values of the adjective as a postpositional nominal modifier (upper part of the graph). Such an arrangement agrees with Greenberg’s proposal. As can be seen, in the red square where there should be no languages, there is only one: Kiche. It is pretty clear how this language seems to be an exception to the Universal since there is a broad white space in the lower right corner without this language. Therefore, the position restriction in that part of the graph by the VSO languages seems clear.

In the graph of the figure 21.38, we have decided to highlight two clusters at the two extremes of the graph’s vertical axis. Although the development of 3 clearly defined clusters and their behavior



Figure 21.37: Subjects and Adjectives (Nominal Modifiers) Direction Checking

we leave for Universal 19, which deals specifically with this phenomenon; we want to highlight the behavior of the adjectives about the direction of the subjects. The group of languages with a subject preceding the verb tends to be located in the lower left corner, thus showing a rather extreme behavior (as Japanese or Bhojpuri). Languages whose preference for subject precedence is not so extreme also show greater flexibility (in fact, relatively proportional to the subject flexibility of each language), as we can see in Gothic, for example. Subsequently, we find some languages in the upper left, although these do not have such an extreme direction and greater flexibility. It could be said that such a central restriction draws a clear crescent, which makes it impossible for languages with flexibility in the preposition of the subject to have rigidity in the direction of the adjective and vice versa. It is also relevant to highlight the more extreme crescent that occurs in the opposite direction on the right side of the graph. However, this is already implicitly included in Greenberg's words. We found occurrences of different types and flexibility in the part of adjectives after the noun they modify. At the same time, there is a much more homogeneous behavior (many overlaps are seen) in the case of adjectives preceding the nouns they modify.

On the other hand, we decided to test the relationship between two elements that usually have a robust syntactic relationship, such as adjectives as nominal modifiers and adverbs as adjectival

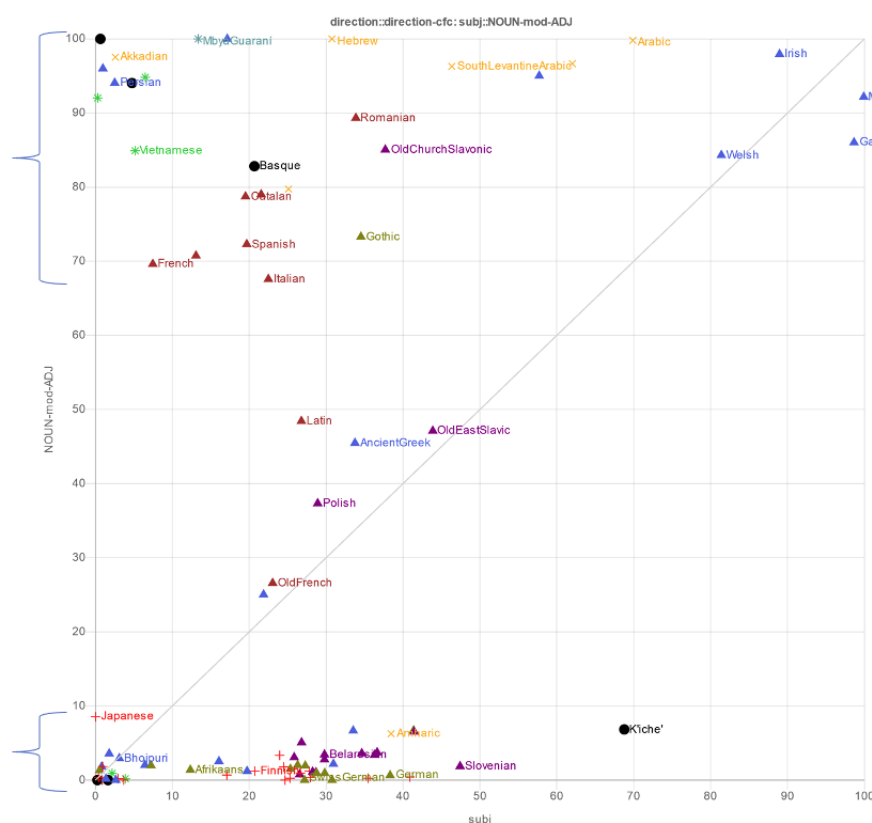


Figure 21.38: Subjects and Adjectives (Nominal Modifiers) Direction Proposal

modifiers, but that was not checked before, which are shown in the figures 21.39 and 21.40.

First, we can see in figure 21.39 how it is possible to formulate a Universal of the type. If a language places adjectives as modifiers before the nouns in the dominant order, it will place adverbs before the adjective they modify. In this case, there is only one exception: Sanskrit. Moreover, such a language appears on the borderline between being an exception or not and should also be quarantined since the texts used in such a language, given the temporal restrictions of the sources, are very marked.

However, what we find even more interesting is the pattern drawn in figure 21.40. We wanted to highlight an inverse “L”, which gives us an interesting picture of the constraints that languages exhibit. In this case, occurrences can be observed along the right-hand side and the bottom of the graph. This means there is an explicit restriction on languages exhibiting the opposite distribution, i.e., high values of adverbs postposed to the adjective with high values of adjectives preceding nouns. We find languages with adjectives preceding nouns, but adverbs always appear before the adjective. We can find languages that place the adverb after the adjective it modifies, but in that case, the language will place the adjectives after the modified noun and not before it.

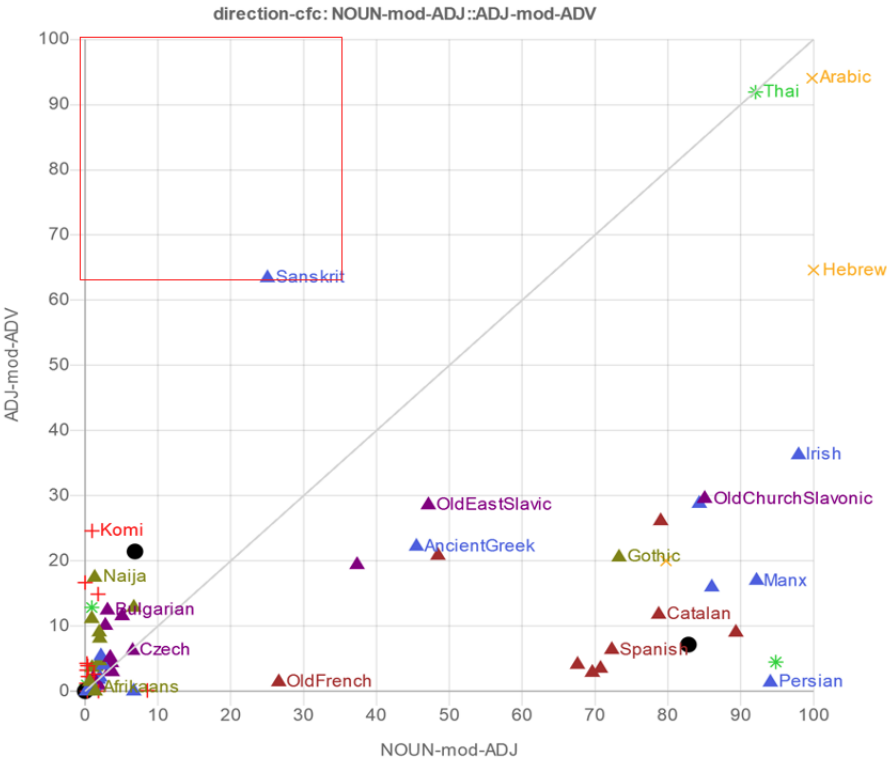


Figure 21.39: Adjectives and Adverbs (Adjectival Modifiers) Direction Checking

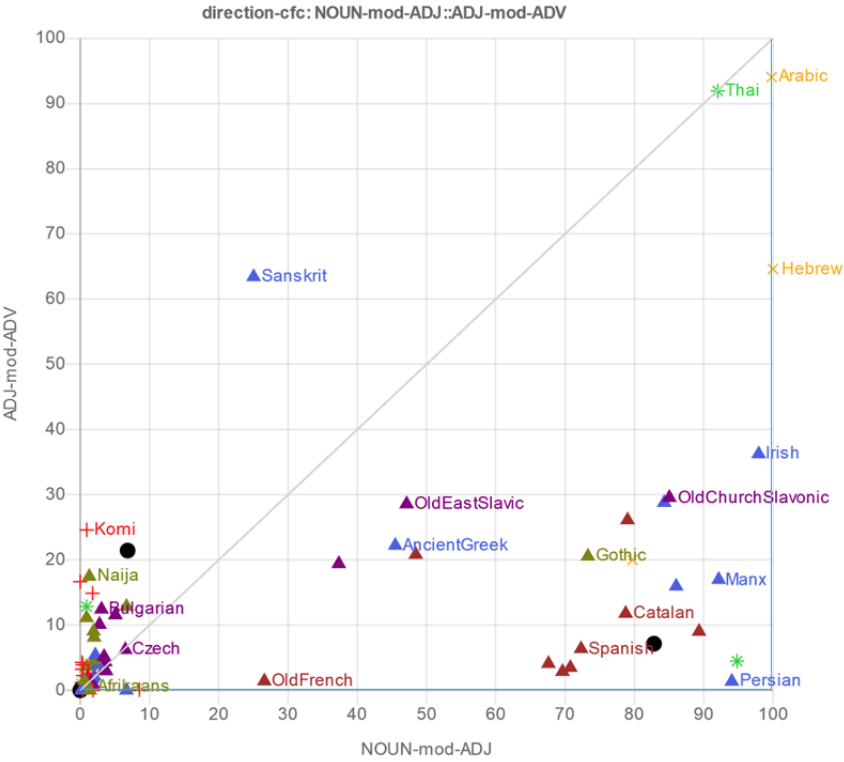


Figure 21.40: Adjectives and Adverbs (Adjectival Modifiers) Direction Proposal

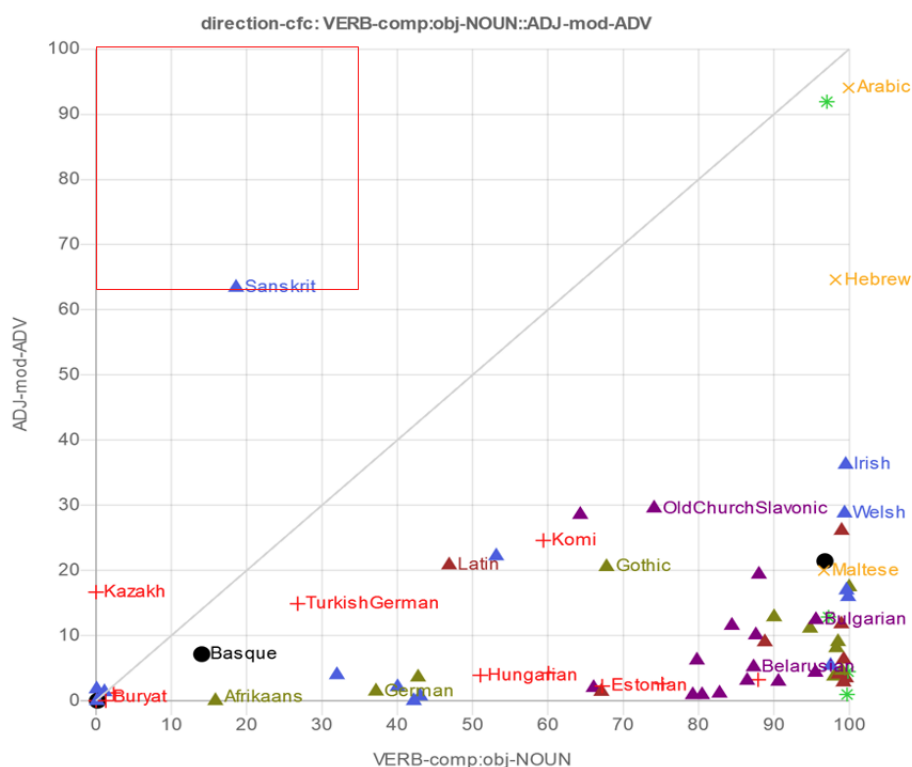


Figure 21.41: Nominal Objects and Adverbs (Adjectival Modifiers) Checking

Similarly, we wanted to check the correlation between the direction of nominal objects and adverbs modifying adjectives; according to researchers like Dryer (1992), the direction of the direct object and the modifying adjective is not a correlative pair. Therefore, a priori, there should be no similar behavior in the graphs 21.41 and 21.42.

As can be seen, first of all, in the figure 21.41, we observe how, again, the languages avoid the upper left corner. We only find Sanskrit right on the border, as explained above. In addition, we think it is very significant to note that no language is found in the vicinity of the red square either; there seems to be a very clear restriction in this corner that does not allow even mixed directions to approach such a position.

In figure 21.42, we see, once again, the same inverted L. Although the distribution of the languages is slightly different (many from the left zone move to the right zone), the pattern drawn is the same. We can even draw a triangle beyond this L, in which only the previously described language (Sanskrit) and Kazakh (being very close to the boundary) escape. In this case, therefore, we can be sure that languages have a Universal tendency to consistently exhibit a higher proportion of nominal direct objects after the verb than of adverbs as modifiers of the postposed adjective. That is, most languages tend to show very low values of occurrences of postposed adverbs. However,

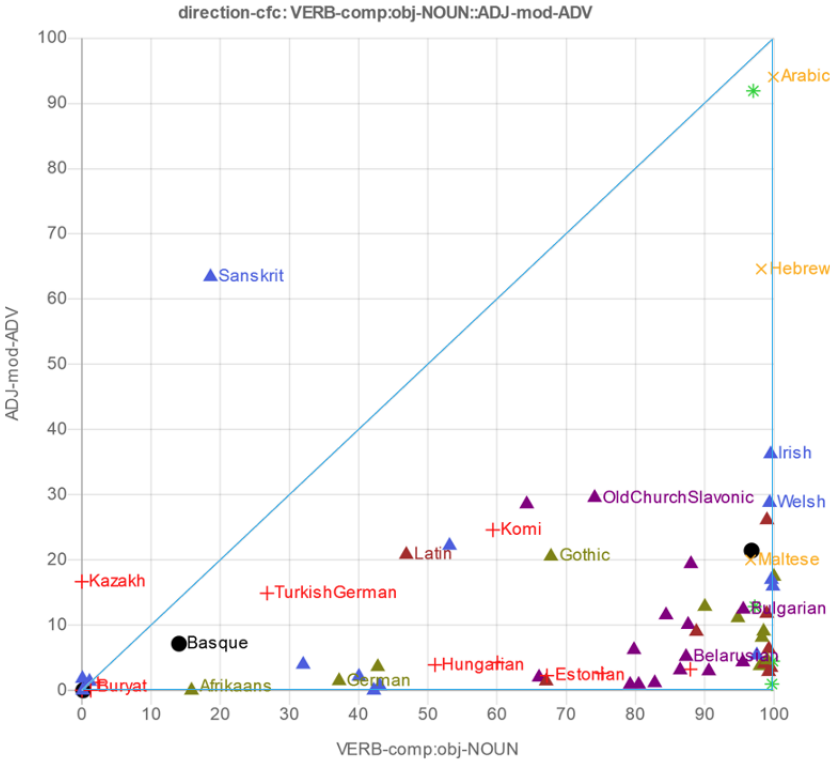


Figure 21.42: Nominal Objects and Adverbs (Adjectival Modifiers) Proposal

those languages that do tend to have a high value (higher than the other parameter) of nominal direct object postposition.

21.17.4 Trichotomous Expression Based on Grew-Match Data

Once we have carried out the analysis of the Universal 17 we can say that we have obtained the following data:

1. Regarding the quantitative check using Grew-Match, we know that the original Universal formulated by Greenberg (applied to the VSO type) is fulfilled in 75% of the cases.
2. If we remove the restriction that the direction of the revised adjectives be of noun modification and extend it to the modification of any parts of speech, we also obtain a 75% compliance rate.
3. Subsequently, we wanted to check the opposite tracking of the Universal in SOV type languages, in congruence with the usual behavior of this linguistic type, something that is fulfilled 79% of the time.

4. Finally, we have decided whether SVO-type languages show a degree of compliance comparable to that of VSO languages (as occurs with other Universals such as the tendency to preposition). In this case, we obtain 34% compliance with the Universal.

This, transferred to a trichotomous expression as the one proposed in this thesis, would show a result as in figure 21.43:

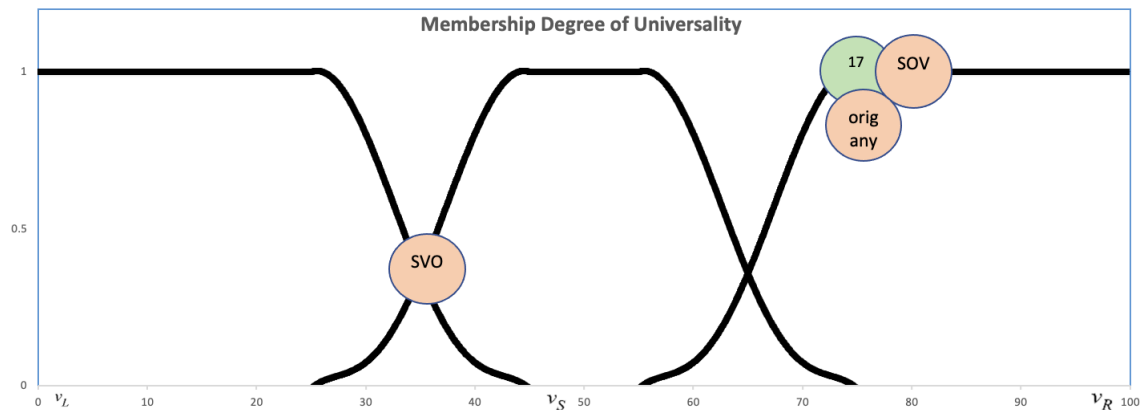


Figure 21.43: Trichotomous Expression for Universal 17

In summary, as can be seen, the Universal proposed by Greenberg is High-Universal and prototypical (although its quantitative values are not close to 100%). The proposal we make to eliminate the restriction that nominal elements are modified also offers values in accordance, and, therefore, we believe that the High-Universal that shows a larger extension and a smaller restriction in its formulation is more valuable. In the case of a comparable proposal that we carried out for SOV languages (adjectives before the modified noun), we can also see how we are dealing with a prototypical High-Universal, and, therefore, the Universal that we formulate is also worth taking into account. On the other hand, the same does not happen with the application to SVO languages that have the adjectives after the modified noun since we are dealing with a Mid-Universal with non-prototypical values close to a Low-Universal. This means that no such correlation can be established in this case. Compared to Dryer's formulation, in which he proposes not to consider the direction of the object to the direction of the adjective, we must indicate that, indeed, through our methodology, we can reach similar conclusions. Moreover, in our case, we can determine that the exceptions come from the SVO-type languages since there is a clear tendency for languages, in general, to place adjectives before nouns in the dominant order. In the case of SOV languages, there are a few cases where the rule is broken; in the case of SVO languages, there is no clear correlation. We can summarise our findings as in table 21.20.

Summary of characteristics of the Universal 17 analysis: 1. Typometrics: The world’s languages tend to place the modifying adjective before the noun and the modifying adverb before the adjective. Alternatively, someplace the adjective after the modifying noun (bottom left of the first pair of graphs). 2. Typometrics: The world’s languages are placed in an inverted L. The free corner is the top one. The free corner is the upper left corner, which Greenberg also indicates. 3. Typometrics: The same inverted L distribution (except for Sanskrit) is observed in the relationship between the direction of the nominal direct object and the direction of the adverb as an adjectival modifier. 4. The original Universal proposed by Greenberg has a fulfillment of 75%. 5. If we analyze the modification of adjectives to any part of speech, we get 75% again. 6. The analysis of SOV languages and their relation to adjectives preceding nouns gives similar results to those of VSO languages (79%). analysis of the SVO languages and the adjectives placed after nouns gives a 34% compliance rate. 7. Greenberg’s original Universal is a High-Universal. 8. The Universal applied to SOV languages is a High-Universal. 9. The original Universal removing the adjective modification to nouns and making it from any part of speech is a High-Universal. 10. The Universal applied to SVO languages is a Mid-Universal, with the potential to be Low-Universal.

Table 21.20: Summary of Characteristics of the Universal 17 Analysis

21.18 Universal 18

UNIVERSAL 18: *When the descriptive adjective precedes the noun, the demonstrative, and the numeral, with overwhelmingly more than chance frequency, do likewise.*

21.18.1 Formalization Proposed for Grew-Match

To check this Universal, we must independently check the order of three elements: adjectives, demonstratives, and numerals. To check the first element, we look at the position of the adjectives in relation to the noun they modify, as indicated by Greenberg in the Universal, (21.191) and

(21.192). For the second element, we look at the position of determiners (and apply the filter that they are demonstratives) and check their position when they determine a noun, (21.193) and (21.194). For the third element, we look at the position of numerals modifying a noun, (21.195) and (21.196).

$$\begin{array}{l} \%18 - ADJ - before \\ pattern\{N[upos = NOUN|PROPN]; N - [mod] - > \\ ADJ; ADJ[upos = ADJ]; ADJ << N\} \end{array} \quad (21.191)$$

(21.191) captures preceding adjectives that modify nouns.

$$\begin{array}{l} \%18 - ADJ - after \\ pattern\{N[upos = NOUN|PROPN]; N - [mod] - > \\ ADJ; ADJ[upos = ADJ]; ADJ >> N\} \end{array} \quad (21.192)$$

(21.192) captures adjectives that modify postpositional nouns.

$$\begin{array}{l} \%18 - DEM - before \\ pattern\{N[upos = NOUN|PROPN]; N - [det] - > \\ DEM; DEM[upos = DET, PronType = Dem]; DEM << N\} \end{array} \quad (21.193)$$

(21.193) captures preceding determiners that determine nouns.

$$\begin{array}{l} \%18 - DEM - after \\ pattern\{N[upos = NOUN|PROPN]; N - [det] - > \\ DEM; DEM[upos = DET, PronType = Dem]; DEM >> N\} \end{array} \quad (21.194)$$

(21.194) captures determiners that determine postpositional nouns.

$$\begin{array}{l} \%18 - NUM - before \\ pattern\{N[upos = NOUN|PROPN]; N - [mod] - > \\ NUM; NUM[upos = NUM]; NUM << N\} \end{array} \quad (21.195)$$

(21.195) captures the numerals before their nouns.

$$\begin{array}{l} \%18 - NUM - after \\ pattern\{N[upos = NOUN|PROPN]; N - [mod] - > \\ NUM; NUM[upos = NUM]; NUM >> N\} \end{array} \quad (21.196)$$

(21.196) captures the numerals after their noun.

The way to check this is first to select the languages with the adjective before the noun. To make a check in categorical terms, we determine that such a characteristic exists if the proportion is higher than 66.66% (according to the dominance calculation of the double previously mentioned). In purely quantitative terms, establishing this type of threshold would not be necessary. Of all those languages, we assigned the adjective value before the noun and checked whether it is also respected to the other two elements. It is interesting to do this check separately for two reasons. First, we obtain a larger number of occurrences. Secondly, we can discriminate which is the exact element that does not contribute to validating the Universal in case it occurs.

However, as mentioned above, we are also interested in testing alternative variables to the exact Universal proposed by Greenberg to determine whether his formulation can be extended or slightly modified. In this case, we believe that it makes sense to analyze elements with behavior similar to the categories enunciated by Greenberg.

First, we want to check what is the behavior of definite articles. This element is particularly relevant as it has traditionally been assigned as a non correlation pair (Dryer, 1992). We propose the formalizations (21.197) and (21.198), where we look for precisely the two possible orders of definite articles determining a name.

$$\begin{array}{l} \%18 - ART - after \\ pattern\{N[upos = NOUN|PROPN]; N - [det] - > ART; ART[upos = \\ DET, PronType = Art]; ART >> N\} \end{array} \quad (21.197)$$

(21.197) captures articles that precede their noun.

$$\begin{array}{l} \%18 - ART - before \\ pattern\{N[upos = NOUN|PROPN]; N - [det] - > ART; ART[upos = \\ DET, PronType = Art]; ART << N\} \end{array} \quad (21.198)$$

(21.198) captures articles postposed to their noun.

Similarly, we can check the behavior of indefinites in the (21.199) and (21.200) patterns, using the same formal structure and modifying the necessary tags to change from definite to indefinite items.

$$\begin{array}{l}
 \%18 - IND - after \\
 pattern\{N[upos = NOUN|PROPN]; N - [det] - > IND; IND[upos = \\
 DET, PronType = Ind]; IND >> N\}
 \end{array} \tag{21.199}$$

(21.199) captures the indefinites that precede its noun.

$$\begin{array}{l}
 \%18 - IND - before \\
 pattern\{N[upos = NOUN|PROPN]; N - [det] - > IND; IND[upos = \\
 DET, PronType = Ind]; IND << N\}
 \end{array} \tag{21.200}$$

(21.200) captures indefinites placed after its noun.

Subsequently, we can check whether the adjective's position is a possible predictor, in general, of any element modifying a noun since this is the most prototypical modifier. For this purpose, we propose the formalizations (21.201) and (21.202). In these we simply seek to find the occurrences of any element, prototypical or not, modifying the noun, differentiated by position.

$$\begin{array}{l}
 \%18 - any - before \\
 pattern\{N[upos = NOUN|PROPN]; N - [mod] - > ANY; \\
 ANY[upos = ADJ|ADV|INTJ|NOUN|PROPN| \\
 VERB|ADP|AUX|CCONJ|DET|NUM|PART|PRON| \\
 SCONJ|SYM]; ANY << N\}
 \end{array} \tag{21.201}$$

(21.201) captures any modifiers that precede their noun.

$$\begin{array}{l}
 \%18 - any - after \\
 pattern\{N[upos = NOUN|PROPN]; N - [mod] - > ANY; \\
 ANY[upos = ADJ|ADV|INTJ|NOUN|PROPN| \\
 VERB|ADP|AUX|CCONJ|DET|NUM|PART|PRON| \\
 SCONJ|SYM]; ANY >> N\}
 \end{array} \tag{21.202}$$

(21.202) captures any modifiers after their noun.

It is equally interesting to look at the behavior of nouns depending on another noun to see if the exact correlation is observed with adjectives. However, in this case, we can find three quite similar possibilities regarding this behavior: a noun modifying another, a noun in apposition to another, and a noun in composition to another. The most relevant case is the first one, which shows a clearer dependency relationship. However, in the other two cases, there is also a dependency

relation in Tesnière’s framework, and both are not at the same level. For example, in the example “the Kraft company”, “Kraft” is a proper noun dependent on a noun (“company”) in apposition.

It should be noted that in the formalizations we propose, we do not contemplate that both are proper nouns. This has been done in order not to obtain false positives of proper nouns expressed in two words that have been mislabeled. Although this problem should not occur, since there is a specific dependency for this (flat), applicable to multiword expressions, we have found some treebanks in which there is no good labeling of this (we suppose that in many cases, it is because of an automatic UD conversion). Therefore, we have taken this decision, which inevitably reduces the number of occurrences that we can find in the languages since there may be other cases that are not captured either because of this. Fortunately, the problem highlighted has an easy solution, as the annotation scheme provides for these situations, and the annotations simply need to be improved or made with greater supervision.

In short, we propose formalizations (21.203) - (21.206).

$$\begin{array}{l} \%18 - n - mod - hbefore \\ pattern\{N[upos = NOUN|PROPN]; N - [mod] - > N2; N2[upos = \\ NOUN]; N2 >> N\} \end{array} \quad (21.203)$$

(21.203) captures a placed after noun modifying another noun.

$$\begin{array}{l} \%18 - n - mod - hafter \\ pattern\{N[upos = NOUN|PROPN]; N - [mod] - > N2; N2[upos = \\ NOUN]; N2 << N\} \end{array} \quad (21.204)$$

(21.204) captures a preceding noun modifying another noun.

$$\begin{array}{l} \%18 - n - appos - hbefore \\ pattern\{N[upos = NOUN|PROPN]; N - [appos] - > N2; N2[upos = \\ NOUN]; N2 >> N\} \end{array} \quad (21.205)$$

(21.205) captures a placed after a noun in apposition with another noun.

$$\begin{array}{l} \%18 - n - appos - hafter \\ pattern\{N[upos = NOUN|PROPN]; N - [appos] - > N2; N2[upos = \\ NOUN]; N2 << N\} \end{array} \quad (21.206)$$

(21.206) captures a preceding noun in apposition with another noun.

In the first two, we look for the two possible orders of a name modifying another name (for example, “tragedie světa” in Czech). In the next two, we look for a name in apposition to another name (for example, “rahkaa kuningatar” in Finnish). We are not interested in differentiating these labels but in checking invariances regarding word order in these three dependencies. That is, to confirm if, for example, a language always places the dependent noun after or before the governing noun, apparently working similarly to adjectives, as we can see in “The Kraft group” vs. “El grupo Inditex” in English and Spanish.

We are also genuinely interested in checking whether the behavior of verbs and auxiliaries as noun modifiers, whatever their type (alone or in a longer clause), is comparable to that of adjectives in the languages of the world. For this, we propose formalizations (21.207), (21.208), (21.209) and (21.210), corresponding to verbs before the modified noun, verbs after the modified noun, auxiliaries before the modified noun, and auxiliaries after the modified noun, respectively.

$$\begin{array}{l} \%18 - vRel - before \\ pattern\{N[upos = NOUN|PROPN]; N - [mod] - > V; V[upos = VERB]; V << N\} \end{array} \quad (21.207)$$

(21.207) captures a preceding verb modifying a noun.

$$\begin{array}{l} \%18 - vRel - after \\ pattern\{N[upos = NOUN|PROPN]; N - [mod] - > V; V[upos = VERB]; V >> N\} \end{array} \quad (21.208)$$

(21.208) captures a placed after verb modifying a noun.

$$\begin{array}{l} \%18 - auxRel - before \\ pattern\{N[upos = NOUN|PROPN]; N - [mod] - > AUX; AUX[upos = AUX]; AUX << N\} \end{array} \quad (21.209)$$

(21.209) captures a preceding auxiliary modifying a noun.

$$\begin{array}{l} \%18 - auxRel - after \\ pattern\{N[upos = NOUN|PROPN]; N - [mod] - > AUX; AUX[upos = AUX]; AUX >> N\} \end{array} \quad (21.210)$$

(21.210) captures a placed after auxiliary modifying a noun.

Just as we have seen the languages that have one noun modifying another where it will often coincide with genitives (either marked by case, as in Polish, for example, or by mere proximity, as in Vietnamese, for example), it is also interesting to discuss these cases when an adposition is used (and to see other cases of noun-dependent adposition as well), as could be “la pau dels valents” in Catalan. Therefore, we propose formalizations (21.211) and (21.212) to capture this.

$$\begin{array}{l}
 \%18 - \text{adp} - \text{before} \\
 \text{pattern}\{N[\text{upos} \quad = \quad \text{NOUN}|\text{PROPN}]; N \quad - \quad [\text{mod}] - \quad > \\
 \text{ADP}; \text{ADP}[\text{upos} = \text{ADP}]; \text{ADP} << N\}
 \end{array} \quad (21.211)$$

(21.211) captures a preceding adposition modifying a noun.

$$\begin{array}{l}
 \%18 - \text{adp} - \text{after} \\
 \text{pattern}\{N[\text{upos} \quad = \quad \text{NOUN}|\text{PROPN}]; N \quad - \quad [\text{mod}] - \quad > \\
 \text{ADP}; \text{ADP}[\text{upos} = \text{ADP}]; \text{ADP} >> N\}
 \end{array} \quad (21.212)$$

(21.212) captures a placed after adposition modifying a noun.

Adverbs are another part of speech that we consider relevant when modifying a noun. We believe that they have enjoyed little prominence in general analyses in a grammar-based perspective, since they are not considered a canonical modification. However, we believe that it is pretty productive and widespread in languages and that we should try to understand whether there is a correspondence with the behavior of adjectives (although we hypothesize that there is not). We are talking about cases such as “Allerede søndag” in Norwegian (Bokmaal). We can formalize these situations as in (21.213) and (21.214).

$$\begin{array}{l}
 \%18 - \text{adv} - \text{before} \\
 \text{pattern}\{N[\text{upos} \quad = \quad \text{NOUN}|\text{PROPN}]; N \quad - \quad [\text{mod}] - \quad > \\
 \text{ADV}; \text{ADV}[\text{upos} = \text{ADV}]; \text{ADV} << N\}
 \end{array} \quad (21.213)$$

(21.213) captures a preceding adverb modifying a noun.

$$\begin{array}{l}
 \%18 - \text{adv} - \text{after} \\
 \text{pattern}\{N[\text{upos} \quad = \quad \text{NOUN}|\text{PROPN}]; N \quad - \quad [\text{mod}] - \quad > \\
 \text{ADV}; \text{ADV}[\text{upos} = \text{ADV}]; \text{ADV} >> N\}
 \end{array} \quad (21.214)$$

(21.214) captures a placed after adverb modifying a noun.

Finally, we can also make an extension opposite to the one related to any part of speech as a modifier. That is, to check whether the order of the adjective modifying the noun is comparable to the order of the adjective modifying any other element less prototypical than the noun or, on the contrary, there is more room for flexibility (or an opposite behavior). For example, an adjective modifying a pronoun, as in “todos juntos” in Spanish. The patterns would be:

%18 – *ADJ* – *any* – *before*

pattern{*ANY*[*upos* = *ADJ*|*ADV*|*INTJ*|*NOUN*|*PROPN*|

VERB|*ADP*|*AUX*|*CCONJ*|*DET*|*NUM*|*PART*|*PRON*|

SCONJ|*SYM*]; *ANY* – [*mod*]– > *ADJ*; *ADJ*[*upos* = *ADJ*]; *ANY* >>

ADJ}

(21.215)

(21.215) captures a preceding adjective modifying any part of speech.

%18 – *ADJ* – *any* – *after*

pattern{*ANY*[*upos* = *ADJ*|*ADV*|*INTJ*|*NOUN*|*PROPN*|

VERB|*ADP*|*AUX*|*CCONJ*|*DET*|*NUM*|*PART*|*PRON*|

SCONJ|*SYM*]; *ANY* – [*mod*]– > *ADJ*; *ADJ*[*upos* = *ADJ*]; *ANY* <<

ADJ}

(21.216)

(21.216) captures a placed after adjective modifying any part of speech.

21.18.2 Trichotomous Expression Based on Grew-Match Data

The Universal 18 includes two consequences simultaneously that must be true if the implication is given for the Universal to be true. That demonstratives precede the noun when adjectives do, individually, occur in 100% of cases. That numerals precede the noun when adjectives do, individually, occurs in 94% of cases. This already establishes a slight difference (not too relevant) that is interesting to note. The fulfillment of the Universal in general is 90%. Although one might think that this should be 94%, we should note that the individual checks may include languages that establish an existing dependency in a language while not attesting to occurrences of the other dependency of the Universal. A language may have numerals but no demonstratives. This fact causes the universe of languages with four elements (adjectives, nouns, demonstratives, and numerals) to be lower than the universe that contemplates only three elements, generating this

slight change in percentages.

In addition, we wanted to check what happens with other types of determiners and their behavior in terms of direction. We obtained 94% of the cases in which a determiner of a definite article precedes the noun in the case where the adjective precedes the noun. We obtained 100% of cases in relation to indefinites preceding nouns. We also wanted to check the consistency of the dependency, in general terms, of any noun-dependent element preceding it in such an implication, with a result of 70%.

These data are subsequently transferred to the trichotomous expression of the figure 21.44.

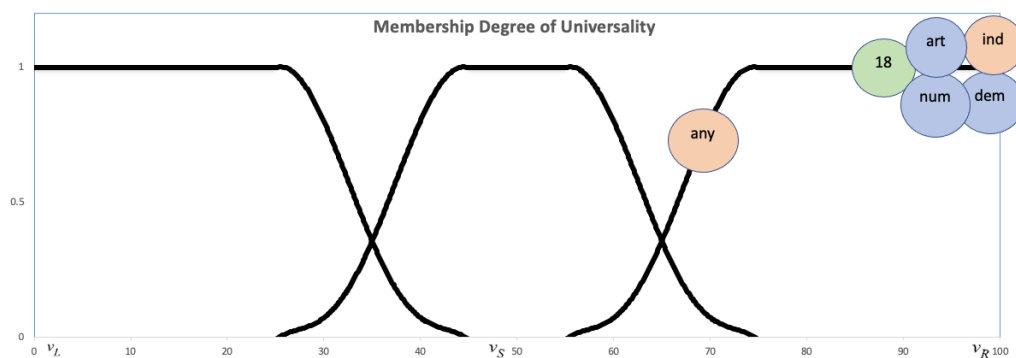


Figure 21.44: Trichotomous Expression for Universal 18

In the case of the description, in terms of consistency of head-finality predicted by the dependence of the adjective as a preceding modifier of the noun, we are dealing with a non-prototypical High-Universal. Therefore, the results such a claim can offer us in Typology are positive, although it is preferable to carry out individual and more specific checks (the rest of the cases). Both the Greenberg Universal in its totality and original formulation and the individual verification of the two parts affected by this Universal, as well as the dependence of the articles and the indefinites, are High-Universal prototypical. In other words, their importance in a typological classification is high. We can summarise our findings as in table 21.21.

Summary of characteristics of the Universal 18 analysis: 1. Greenberg’s original Universal is fulfilled in 90% of the cases. 2. The final result of the Universal is the sum of the fulfillment of 100% of the precedence of demonstratives when adjectives precede and 94% of the precedence of numerals when adjectives precede. 3. The compliance of preceding articles when adjectives precede is 94%. 4. The compliance of the indefinite preceding when adjectives precede is 100%. 5. The compliance of this Universal applied to the harmony of modifying an adjective to any part of speech is 70%. 6. All the cases we have reviewed, including Greenberg’s Universal or our proposed extension to indefinites, are High-Universals.

Table 21.21: Summary of Characteristics of the Universal 18 Analysis

21.19 Universal 19

UNIVERSAL 19: *When the general rule is that the descriptive adjective follows, there may be a minority of adjectives that usually precede, but when the general rule is that descriptive adjectives precede, there are no exceptions.*

21.19.1 How to Check It in Typometrics

Technical Decisions: 1. Firstly, the Universal is analyzed with Typometrics (Gerdes et al., 2021). 2. We select the SUD annotation scheme because we can get results more comparable to the classical typological claims. 3. We select Graph 1D since this Universal only mentions adjectives’ behavior 4. We set a filter of utterances to avoid misleading data. 5. We select <i>direction – cfc</i> in the only axis and check for <i>NOUN – mod – ADJ</i> since we want to check the canonical relation of adjectives modifying nouns.

Table 21.22: Technical Decisions for Checking Universal 19 in Typometrics

21.19.2 Formalization Proposed for Grew-Match

If we want to analyze by Grew-Match this same Universal proposed by Greenberg, we simply need to create two elements, which we can call *N* and *ADJ*. We assign the *N* tag the value of [*upos* = *NOUN*] and the *ADJ* tag the value of [*upos* = *ADJ*]. Then, we ask for an adjective dependency by modifying a name as a head, i.e., *N* – [*mod*]– > *ADJ*. Finally, we present a different order in each of the two patterns: *N* << *ADJ* or *N* >> *ADJ*. That is, the complete patterns would be like in Universal 17 (21.17): (21.217) and (21.218).

$$\begin{array}{l} \%19 - ADJ - befN \\ pattern\{N[upos = NOUN]; N - [mod] - > ADJ; ADJ[upos = \\ ADJ]; ADJ << N\} \end{array} \quad (21.217)$$

(21.217) captures preceding adjectives that modify nouns.

$$\begin{array}{l} \%19 - ADJ - aftN \\ pattern\{N[upos = NOUN]; N - [mod] - > \\ ADJ; ADJ[upos = ADJ]; ADJ >> N\} \end{array} \quad (21.218)$$

(21.218) captures placed after adjectives that modify nouns.

However, as we have done with other Universals, we can go beyond what Greenberg proposed and work on slightly different situations to try to extend the Universal, nuance it, or directly propose new Universals. We think that the most useful may be to check other parts of speech that are typically related to adjectives as modifiers (although their relation is not as prototypical as that of the noun): *NUM* (behaving similar to a pronoun), *PRON* or *PROPN*. It is also possible to check the relation of verbs and adjectives as predicative, but, as we mentioned before, it is a more complex dependency where more agents are involved. Then, we prefer not to comment deeply on any result, only to propose a theoretical formalization. We can also check the data in a more general and extended way, i.e., checking when adjectives modify any part of speech or when any part of speech modifies a noun, the opposite version of the Universal.

The patterns included in these proposals are (21.219) - (21.230).

$$\begin{array}{l} \%19 - ADJ - aftNUM \\ pattern\{NUM[upos = NUM]; NUM - [mod] - > ADJ; ADJ[upos = \\ ADJ]; ADJ >> NUM\} \end{array} \quad (21.219)$$

(21.219) captures placed after adjectives that modify numerals.

$$\begin{array}{l} \%19 - ADJ - bef NUM \\ pattern\{NUM[upos = NUM]; NUM - [mod] - > ADJ; ADJ[upos = \\ ADJ]; ADJ << NUM\} \end{array} \quad (21.220)$$

(21.220) captures preceding adjectives that modify numerals.

$$\begin{array}{l} \%19 - ADJ - bef PRON \\ pattern\{PRON[upos = PRON]; PRON - [mod] - > \\ ADJ; ADJ[upos = ADJ]; ADJ << PRON\} \end{array} \quad (21.221)$$

(21.221) captures preceding adjectives that modify pronouns.

$$\begin{array}{l} \%19 - ADJ - aft PRON \\ pattern\{PRON[upos = PRON]; PRON - [mod] - > \\ ADJ; ADJ[upos = ADJ]; ADJ >> PRON\} \end{array} \quad (21.222)$$

(21.222) captures placed after adjectives that modify pronouns.

$$\begin{array}{l} \%19 - ADJ - aft PROP N \\ pattern\{PROP N[upos = PROP N]; PROP N - [mod] - > \\ ADJ; ADJ[upos = ADJ]; ADJ >> PROP N\} \end{array} \quad (21.223)$$

(21.223) captures placed after adjectives that modify proper nouns.

$$\begin{array}{l} \%19 - ADJ - bef PROP N \\ pattern\{PROP N[upos = PROP N]; PROP N - [mod] - > \\ ADJ; ADJ[upos = ADJ]; ADJ << PROP N\} \end{array} \quad (21.224)$$

(21.224) captures preceding adjectives that modify proper nouns.

$$\begin{array}{l} \%19 - ADJ - bef VERB \\ pattern\{VERB[upos = VERB]; VERB - [comp : pred] - > \\ ADJ; ADJ[upos = ADJ]; ADJ << VERB\} \end{array} \quad (21.225)$$

(21.225) captures preceding adjectives as predicatives.

$$\begin{array}{l}
\%19 - ADJ - aft VERB \\
pattern\{VERB[upos = VERB]; VERB - [comp : pred] - > \\
ADJ; ADJ[upos = ADJ]; ADJ >> VERB\}
\end{array} \tag{21.226}$$

(21.226) captures placed after adjectives as predicatives.

$$\begin{array}{l}
\%19 - ANY - bef N \\
pattern\{N[upos = NOUN]; N - [mod] - > ANY; \\
ANY[upos = ADJ|ADV|INTJ|NOUN|PROPN|VERB| \\
ADP|AUX|CCONJ|DET|NUM|PART|PRON|SCONJ| \\
SYM]; ANY << N\}
\end{array} \tag{21.227}$$

(21.227) captures any part of speech preceding and modifying a noun.

$$\begin{array}{l}
\%19 - ANY - aft N \\
pattern\{N[upos = NOUN]; N - [mod] - > ANY; \\
ANY[upos = ADJ|ADV|INTJ|NOUN|PROPN|VERB| \\
ADP|AUX|CCONJ|DET|NUM|PART|PRON|SCONJ| \\
SYM]; ANY >> N\}
\end{array} \tag{21.228}$$

(21.228) captures any part of speech placed after and modifying a noun.

$$\begin{array}{l}
\%19 - ADJ - bef ANY \\
pattern\{ADJ[upos = ADJ]; ANY - [mod] - > ADJ; \\
ANY[upos = ADJ|ADV|INTJ|NOUN|PROPN|VERB| \\
ADP|AUX|CCONJ|DET|NUM|PART|PRON|SCONJ| \\
SYM]; ANY >> ADJ\}
\end{array} \tag{21.229}$$

(21.229) captures preceding adjectives modifying any part of speech.

$$\begin{array}{l}
\%19 - ADJ - aft ANY \\
pattern\{ADJ[upos = ADJ]; ANY - [mod] - > ADJ; \\
ANY[upos = ADJ|ADV|INTJ|NOUN|PROPN|VERB| \\
ADP|AUX|CCONJ|DET|NUM|PART|PRON|SCONJ| \\
SYM]; ANY << ADJ\}
\end{array} \tag{21.230}$$

(21.230) captures placed after adjectives modifying any part of speech.

In the first two patterns (21.219) and (21.220), we look for the dependence of a numeral modified by an adjective $NUM - [mod] - > ADJ$, with the two possible orders. In the following two patterns (21.221) and (21.222), we look for the dependency of a pronoun modified by an adjective $PRON - [mod] - > ADJ$, with the two possible orders. In the following two patterns (21.223) and (21.224), we formalize the dependency of a proper noun modified by an adjective $PROP N - [mod] - > ADJ$, with the two possible orders, since we have chosen to separate in this case common from proper nouns to check if the adjective position is the same or different. In the following two patterns (21.225) and (21.226), we propose a formalization for the particular case of adjectives as predicatives of a verb, with its two orders ($V - [comp : pred] - > ADJ$), which we will not extend. The following two patterns (21.227) and (21.228) correspond to the modification of a noun by any kind of part of speech, with the respective two orders, i.e., $N - [mod] - > ANY$. The last two patterns (21.229) and (21.230) refer to the opposite case, an adjective being the modifier of any part of speech in general with its two orders: $ANY - [mod] - > ADJ$.

21.19.3 Graphs from Typometrics

Given its simplicity, this Universal can be quickly reviewed in figures 21.45 and 21.46.

Correctly fulfilling this Universal peculiarity would allow a flexible arrangement of languages that tend to place the nominal modifier adjective after the noun (upper part). In other words, these languages may present cases of adjectives before the noun. In contrast, all languages that tend to place the adjective before the noun have no occurrences, according to Greenberg, meaning that they should all be at the 0 end, showing no adjectives after the noun. As can be seen quite clearly, this is not the case.

In the upper part of the axis, we see a reasonably balanced flexibility across the whole spectrum of dominance of after adjectives. In the lower part, which presents the restriction, we see that there are many languages in different percentages ranging from 0 to 10 practically. Therefore, Greenberg's Universal, following his terms in an orthodox way, is not fulfilled. However, we cannot ignore the clearly different behavior in the two groups. The flexibility to manifest occurrences in the opposite direction is much more limited in the adjectives preceding the noun. Thus, Greenberg seems to unravel a rather interesting trend in the world's languages. We believe, therefore, that what should be done with this Universal, rather than discarding it, is to reformulate it so that it is less categorical: instead of stating that "there are no exceptions", we should indicate that "the flexibility is highly limited".

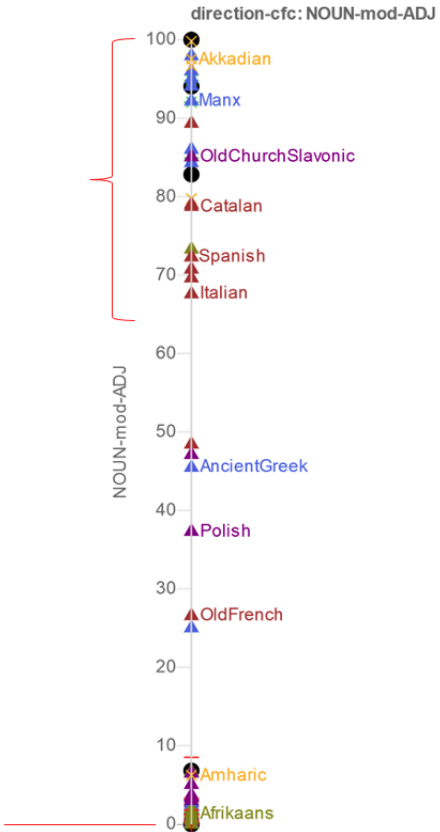


Figure 21.45: Adjectives Modifying Nouns Direction Checking

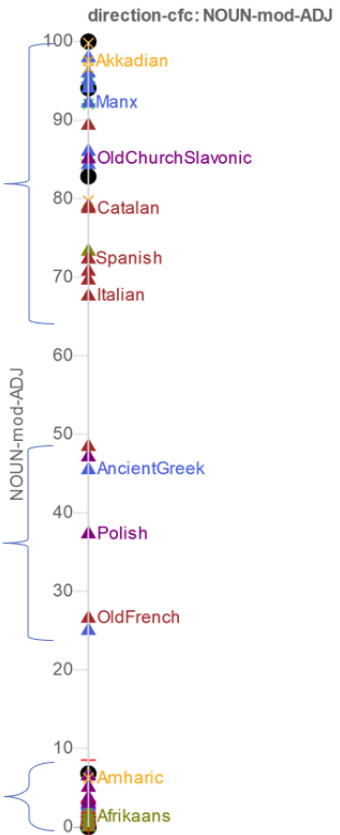


Figure 21.46: Adjectives Modifying Nouns Direction Proposal

In figure 21.46, we show how we can present three different clusters with very interesting characteristics. We have the cluster of the linguistic type of the preceding adjective, the after adjective, and the nondominant order.

1. At the top, we have the languages with adjectives after the noun, which are quite a few and show up at all points in this cluster, with Romance languages showing the most flexibility with direction.
2. Then we have a fairly large and obvious jump without languages, i.e., there is no gradation, which allows us to determine the three groups.
3. Then we find the nondominant order languages. Of this group, we can point out that there are few of them, many of them are usually dead languages, and, in addition, that this group is not located around 50, but before 50. This means, added to the behavior of the previous group and the subsequent group, that there is a clear preference of the languages to put the adjective before the adjective.
4. After another noticeable gap between groups, we find the languages preceding the noun being also many and much closer to the end of the axis (thus allowing fewer occurrences in the opposite direction).

On the other hand, we can show this same axis by reviewing the behavior of adjectives as pronoun modifiers, as shown in figure 21.47.

In this case, we can see how the behavior is different. The group of adjectives before the noun, when we review the pronouns, practically disappears. The central group, with a tendency to be before 50, is now divided in two, with a small group of languages after 55. Most languages appear in the group of adjectives after the pronoun. In general, we see that in this case, the restriction is the opposite: adjectives prefer to be placed after the pronoun, regardless of whether the language tends to place the adjective prenominal or postnominal.

The last related case in the graph has proper nouns as the head of the adjectival modification dependency. These data allow us to see if there are differences in this Typology of nouns, which is not usually considered in grammars and, therefore, is little known information, as shown in figure 21.48.

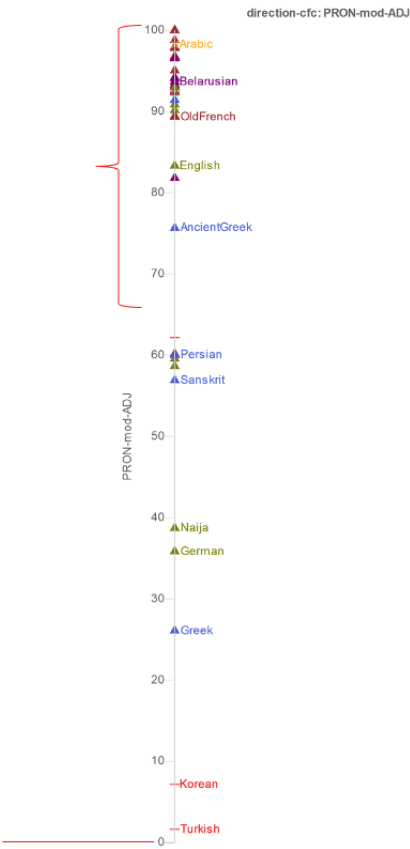


Figure 21.47: Adjectives Modifying Pronouns Direction Checking

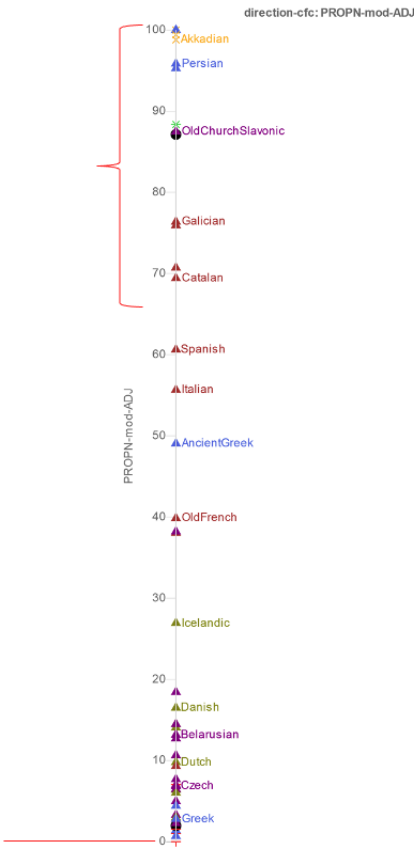


Figure 21.48: Adjectives Modifying Proper Nouns Direction Checking

In this case, a new distribution of languages can be seen, without being similar to the two previous graphs, although with a behavior more similar to common nouns than to pronouns, as was to be expected. In this case, most languages are located in the lower part of the axis; they tend to place adjectives before the proper name. One of the main differences between Greenberg's original and general proposal is that, in this case, there is greater flexibility in such languages, going from a limitation of the 10-group to a 20-group. On the other hand, the central group is not represented in such a clear way since it is more space on the axis and, in addition, it is found in values between 30 and 60, more or less, that is, without the tendency to be located before 50 as in the previous case. Finally, in the group of adjectives after the proper noun, we find fewer languages than in the initial check and very widely spaced, which shows us, in general, the preference for the preceding adjective, even more so than in figure 21.45.

21.19.4 Trichotomous Expression Based on Grew-Match Data

We can sum up the results we observe as follows:

1. The quantitative data on the fulfillment of the Universal show 26% of languages that support it, while 74% of languages are a counterexample.
2. As we have commented above, the main problem with this Universal is not in its prediction, which we have shown to exist, but in its formulation, which we believe to be inaccurate because it is excessively restrictive. If we relax in quantitative terms the prediction and allow languages that place the adjective before the common noun to have a 2% flexibility or occurrences in which adjectives can appear after the noun, the fulfillment becomes 62%.
3. We also wanted to check whether the adjective's position as a nominal modifier (without occurrences in the opposite direction when it precedes) is maintained in other types of dependencies, such as the predicative. In this case, we obtain 48% of cases where this restriction is fulfilled and 52% where it is not. Therefore, there is no Universal trend in this direction, although we obtain much more robust results than when analyzing the direction of the modification dependency. It should not be forgotten that we are being picky about not allowing occurrences in the opposite direction in such a case.

The trichotomous model can represent these data as shown in figure 21.49.

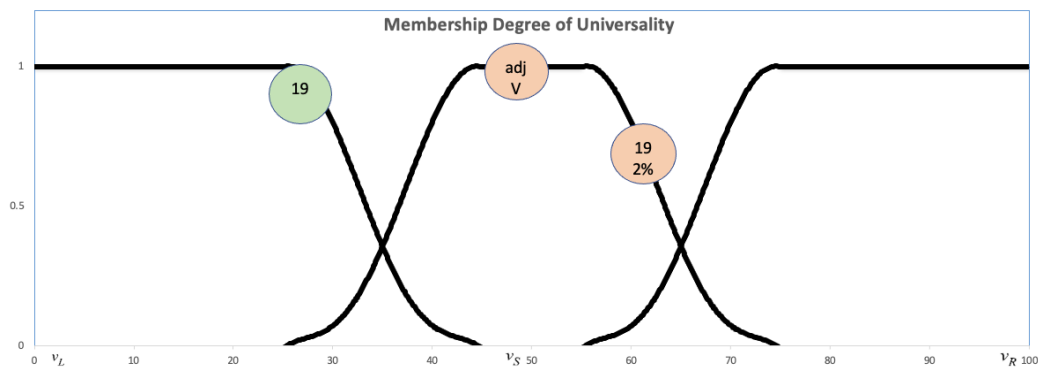


Figure 21.49: Trichotomous Expression for Universal 19

As can be seen, none of the three proposals is a High-Universal and, therefore, cannot be taken as relevant in the terms made here. Greenberg’s original results are in a Low-Universal, while our proposals are in a Mid-Universal. However, we have seen that there is indeed a tendency to be more restrictive in the order of adjectives preceding the noun and a tendency for languages to do so. Therefore, as we have already commented, this Universal cannot be given as valid for Typology. However, we must bear in mind that a reformulation of the Universal in a more general way, as we propose above, would obtain fulfillment of 100% of the cases. However, this formulation is less precise and, therefore, less representative, so we do not propose it so strongly. We can summarise our findings as in table 21.23.

Summary of characteristics of the Universal 19 analysis:
1. Typometrics: There are three groups of languages regarding the direction of the adjective modifying a noun.
2. Typometrics: Most languages place the adjective before the modified noun.
3. Typometrics: Languages that place the adjective before are stricter when allowing occurrences in the opposite direction compared to languages that place the adjective after the modified noun.
4. The original Universal is fulfilled in 26% of the cases. The main cause is that Greenberg considers no flexibility of opposite occurrences.
5. Allowing little flexibility in occurrences of adjectives placed after the result grows.
6. There are no better tendencies when formulating this Universal differently: with another part of speech or kinds of dependencies.
7. The original Universal is a Low-Universal.

Table 21.23: Summary of Characteristics of the Universal 19 Analysis

21.20 Universal 20

UNIVERSAL 20: *When any or all items (demonstrative, numeral, and descriptive adjective) precede the noun, they are always found in that order. If they follow, the order is either the same or opposite.*

21.20.1 Formalization Proposed for Grew-Match

This Universal uses the same elements from Universal 18. However, we cannot use the previously extracted data since Greenberg is now interested in cooccurrence when these elements appear together in an utterance. We must then check the situations where demonstratives, numerals, and adjectives appear together (or at least 2 of them). It is also interesting to check what happens when some appear before and some after, something not mentioned in the Universal. Thus, we want to show specific results in any possible order combination. Some of them are stated as frequent by Greenberg, others are unattested according to him, and he does not check some of them.

Thus, the logical options are:

1. DEM-NUM-ADJ-NOUN
2. DEM-ADJ-NUM-NOUN
3. NUM-DEM-ADJ-NOUN
4. NUM-ADJ-DEM-NOUN
5. ADJ-NUM-DEM-NOUN
6. ADJ-DEM-NUM-NOUN
7. NOUN-DEM-NUM-ADJ
8. NOUN-DEM-ADJ-NUM
9. NOUN-NUM-DEM-ADJ
10. NOUN-NUM-ADJ-DEM
11. NOUN-ADJ-NUM-DEM
12. NOUN-ADJ-DEM-NUM

13. DEM-NOUN-ADJ-NUM
14. DEM-NOUN-NUM-ADJ
15. DEM-ADJ-NOUN-NUM
16. DEM-NUM-NOUN-ADJ
17. NUM-NOUN-DEM-ADJ
18. NUM-NOUN-ADJ-DEM
19. NUM-DEM-NOUN-ADJ
20. NUM-ADJ-NOUN-DEM
21. ADJ-NOUN-DEM-NUM
22. ADJ-NOUN-NUM-DEM
23. ADJ-DEM-NOUN-NUM
24. ADJ-NUM-NOUN-DEM

Greenberg points out that only 1, 7, and 11 are possible (“these four red laptops”, for example). When developing the Universal, he also states that case number 7 is less common than 11, but he does not offer any specific data regarding this. We do not have any data for all the cases involving combinations with both before and after the noun. Dryer (2019) has worked recently in this Universal, offering data from 576 languages based on grammars. It could be interesting to compare the obtained results, which also can show the distance between a grammar-based approach (with known benefits and limitations) and the computational-corpus-based approach (with its own benefits and limitations too). We have to activate permanently the same structure, changing in each case the order constraint using << or >> between the elements. The structure is based on a noun as head and then modifications or determinations of this noun. Given the specificity of the analyzed structure, we do not propose less prototypical alternative constructions that Greenberg has not taken into account. As we have commented, this concrete and complex structure is not well represented in real texts. Therefore, it makes no sense to look for less prototypical structures that, although they might be technically possible, do not appear as occurrences.

For the logical options mentioned before, we can create the following patterns (21.231) - (21.254):

$$\begin{array}{l}
\%20 - 3 - 1 - DemNumAdjNoun \\
pattern\{N[upos = NOUN]; N - [det]- > DEM; DEM[upos = \\
DET, PronType = Dem]; N - [mod]- > NUM; NUM[upos = \\
NUM]; N - [mod]- > ADJ; ADJ[upos = ADJ]; DEM << \\
N; NUM << N; ADJ << N; DEM << NUM; NUM << ADJ\}
\end{array} \tag{21.231}$$

(21.231) captures the order of placing first demonstratives, then numerals, then adjectives, and, finally, nouns.

$$\begin{array}{l}
\%20 - 3 - 2 - DemAdjNumNoun \\
pattern\{N[upos = NOUN]; N - [det]- > DEM; DEM[upos = \\
DET, PronType = Dem]; N - [mod]- > NUM; NUM[upos = \\
NUM]; N - [mod]- > ADJ; ADJ[upos = ADJ]; DEM << \\
N; NUM << N; ADJ << N; DEM << ADJ; ADJ << NUM\}
\end{array} \tag{21.232}$$

(21.232) captures the order of placing first demonstratives, then adjectives, then numerals, and, finally, nouns.

$$\begin{array}{l}
\%20 - 3 - 3 - NumDemAdjNoun \\
pattern\{N[upos = NOUN]; N - [det]- > DEM; DEM[upos = \\
DET, PronType = Dem]; N - [mod]- > NUM; NUM[upos = \\
NUM]; N - [mod]- > ADJ; ADJ[upos = ADJ]; DEM << \\
N; NUM << N; ADJ << N; NUM << DEM; DEM << ADJ\}
\end{array} \tag{21.233}$$

(21.233) captures the order of placing first numerals, demonstratives, adjectives, and nouns.

$$\begin{array}{l}
\%20 - 3 - 4 - NumAdjDemNoun \\
pattern\{N[upos = NOUN]; N - [det]- > DEM; DEM[upos = DET, PronType = Dem]; N - [mod]- > \\
NUM; NUM[upos = NUM]; N - [mod]- > ADJ; ADJ[upos = \\
ADJ]; DEM << N; NUM << N; ADJ << N; NUM << \\
ADJ; ADJ << DEM\}
\end{array} \tag{21.234}$$

(21.234) captures the order of placing first numerals, adjectives, demonstratives, and nouns.

$$\begin{aligned} \%20 - 3 - 5 - AdjNumDemNoun \text{ pattern}\{N[upos = NOUN]; N - \\ [det] - > DEM; DEM[upos = DET, PronType = Dem]; N - [mod] - > \\ NUM; NUM[upos = NUM]; N - [mod] - > ADJ; ADJ[upos = \\ ADJ]; NUM << N; ADJ << N; ADJ << NUM; NUM << DEM\} \end{aligned} \quad (21.235)$$

(21.235) captures the order of placing first adjectives, then numerals, then demonstratives, and, finally, nouns.

$$\begin{aligned} \%20 - 3 - 6 - AdjDemNumNoun \text{ pattern}\{N[upos = NOUN]; N - \\ [det] - > DEM; DEM[upos = DET, PronType = Dem]; N - [mod] - > \\ NUM; NUM[upos = NUM]; N - [mod] - > ADJ; ADJ[upos = \\ ADJ]; DEM << N; NUM << N; ADJ << N; ADJ << \\ DEM; DEM << NUM\} \end{aligned} \quad (21.236)$$

(21.236) captures the order of placing first adjectives, demonstratives, numerals, and nouns.

$$\begin{aligned} \%20 - 3 - 7 - NounDemNumAdj \text{ pattern}\{N[upos = NOUN]; N - \\ [det] - > DEM; DEM[upos = DET, PronType = Dem]; N - [mod] - > \\ NUM; NUM[upos = NUM]; N - [mod] - > ADJ; ADJ[upos = \\ ADJ]; DEM >> N; NUM >> N; ADJ >> N; DEM << \\ NUM; NUM << ADJ\} \end{aligned} \quad (21.237)$$

(21.237) captures the order of placing first nouns, then demonstratives, then numerals, and finally, adjectives.

$$\begin{aligned} \%20 - 3 - 8 - NounDemAdjNum \text{ pattern}\{N[upos = NOUN]; N - \\ [det] - > DEM; DEM[upos = DET, PronType = Dem]; N - [mod] - > \\ NUM; NUM[upos = NUM]; N - [mod] - > ADJ; ADJ[upos = \\ ADJ]; DEM >> N; NUM >> N; ADJ >> N; DEM << \\ ADJ; ADJ << NUM\} \end{aligned} \quad (21.238)$$

(21.238) captures the order of placing first nouns, demonstratives, adjectives, and nouns.

$$\begin{aligned} \%20 - 3 - 9 - NounNumDemAdj \text{ pattern}\{N[upos = NOUN]; N - \\ [det] - > DEM; DEM[upos = DET, PronType = Dem]; N - [mod] - > \\ NUM; NUM[upos = NUM]; N - [mod] - > ADJ; ADJ[upos = \\ ADJ]; DEM >> N; NUM >> N; ADJ >> N; NUM << \\ DEM; DEM << ADJ\} \end{aligned} \quad (21.239)$$

(21.239) captures the order of placing first nouns, then numerals, demonstratives, and adjectives.

$$\begin{aligned} \%20 - 3 - 10 - NounNumAdjDem \text{ pattern}\{N[upos = NOUN]; N - \\ [det] - > DEM; DEM[upos = DET, PronType = Dem]; N - [mod] - > \\ NUM; NUM[upos = NUM]; N - [mod] - > ADJ; ADJ[upos = \\ ADJ]; DEM >> N; NUM >> N; ADJ >> N; NUM << \\ ADJ; ADJ << DEM\} \end{aligned} \quad (21.240)$$

(21.240) captures the order of placing first nouns, then numerals, then adjectives, and, finally, demonstratives.

$$\begin{aligned} \%20 - 3 - 11 - NounAdjNumDem \text{ pattern}\{N[upos = NOUN]; N - \\ [det] - > DEM; DEM[upos = DET, PronType = Dem]; N - [mod] - > \\ NUM; NUM[upos = NUM]; N - [mod] - > ADJ; ADJ[upos = \\ ADJ]; DEM >> N; NUM >> N; ADJ >> N; ADJ << \\ NUM; NUM << DEM\} \end{aligned} \quad (21.241)$$

(21.241) captures the order of placing first nouns, then adjectives, then numerals, and, finally, demonstratives.

$$\begin{aligned} \%20 - 3 - 12 - NounAdjDemNum \text{ pattern}\{N[upos = NOUN]; N - \\ [det] - > DEM; DEM[upos = DET, PronType = Dem]; N - [mod] - > \\ NUM; NUM[upos = NUM]; N - [mod] - > ADJ; ADJ[upos = \\ ADJ]; DEM >> N; NUM >> N; ADJ >> N; ADJ << \\ DEM; DEM << NUM\} \end{aligned} \quad (21.242)$$

(21.242) captures the order of placing first nouns, adjectives, demonstratives, and numerals.

$$\begin{aligned} \%20 - 3 - 13 - \textit{DemNounAdjNum} \textit{ pattern}\{N[\textit{upos} = \textit{NOUN}]; N - \\ [det] - > \textit{DEM}; \textit{DEM}[\textit{upos} = \textit{DET}, \textit{PronType} = \textit{Dem}]; N - [mod] - > \\ \textit{NUM}; \textit{NUM}[\textit{upos} = \textit{NUM}]; N - [mod] - > \textit{ADJ}; \textit{ADJ}[\textit{upos} = \\ \textit{ADJ}]; \textit{DEM} << N; \textit{NUM} >> N; \textit{ADJ} >> N; \textit{ADJ} << \textit{NUM}\} \end{aligned} \quad (21.243)$$

(21.243) captures the order of placing first demonstratives, then nouns, then adjectives, and, finally, numerals.

$$\begin{aligned} \%20 - 3 - 14 - \textit{DemNounNumAdj} \textit{ pattern}\{N[\textit{upos} = \textit{NOUN}]; N - \\ [det] - > \textit{DEM}; \textit{DEM}[\textit{upos} = \textit{DET}, \textit{PronType} = \textit{Dem}]; N - [mod] - > \\ \textit{NUM}; \textit{NUM}[\textit{upos} = \textit{NUM}]; N - [mod] - > \textit{ADJ}; \textit{ADJ}[\textit{upos} = \\ \textit{ADJ}]; \textit{DEM} << N; \textit{NUM} >> N; \textit{ADJ} >> N; \textit{NUM} << \textit{ADJ}\} \end{aligned} \quad (21.244)$$

(21.244) captures the order of placing first demonstratives, then nouns, then numerals, and finally, adjectives.

$$\begin{aligned} \%20 - 3 - 15 - \textit{DemAdjNounNum} \textit{ pattern}\{N[\textit{upos} = \textit{NOUN}]; N - \\ [det] - > \textit{DEM}; \textit{DEM}[\textit{upos} = \textit{DET}, \textit{PronType} = \textit{Dem}]; N - [mod] - > \\ \textit{NUM}; \textit{NUM}[\textit{upos} = \textit{NUM}]; N - [mod] - > \textit{ADJ}; \textit{ADJ}[\textit{upos} = \\ \textit{ADJ}]; \textit{DEM} << N; \textit{NUM} >> N; \textit{ADJ} << N; \textit{DEM} << \textit{ADJ}\} \end{aligned} \quad (21.245)$$

(21.245) captures the order of placing first demonstratives, then adjectives, then nouns, and, finally, numerals.

$$\begin{aligned} \%20 - 3 - 16 - \textit{DemNumNounAdj} \textit{ pattern}\{N[\textit{upos} = \textit{NOUN}]; N - \\ [det] - > \textit{DEM}; \textit{DEM}[\textit{upos} = \textit{DET}, \textit{PronType} = \textit{Dem}]; N - [mod] - > \\ \textit{NUM}; \textit{NUM}[\textit{upos} = \textit{NUM}]; N - [mod] - > \textit{ADJ}; \textit{ADJ}[\textit{upos} = \\ \textit{ADJ}]; \textit{DEM} << N; \textit{NUM} << N; \textit{ADJ} >> N; \textit{DEM} << \textit{NUM}\} \end{aligned} \quad (21.246)$$

(21.246) captures the order of placing first demonstratives, then numerals, then nouns, and finally, adjectives.

$$\begin{aligned} \%20 - 3 - 17 - \textit{NumNounDemAdj} \textit{ pattern}\{N[\textit{upos} = \textit{NOUN}]; N - \\ [det] - > \textit{DEM}; \textit{DEM}[\textit{upos} = \textit{DET}, \textit{PronType} = \textit{Dem}]; N - [mod] - > \\ \textit{NUM}; \textit{NUM}[\textit{upos} = \textit{NUM}]; N - [mod] - > \textit{ADJ}; \textit{ADJ}[\textit{upos} = \\ \textit{ADJ}]; \textit{DEM} >> N; \textit{NUM} << N; \textit{ADJ} >> N; \textit{DEM} << \textit{ADJ}\} \end{aligned} \quad (21.247)$$

(21.247) captures the order of placing first numerals, then nouns, then demonstratives, and, finally, adjectives.

$$\begin{array}{l} \%20 - 3 - 18 - NumNounAdjDem \text{ pattern}\{N[upos = NOUN]; N - \\ [det] - > DEM; DEM[upos = DET, PronType = Dem]; N - [mod] - > \\ NUM; NUM[upos = NUM]; N - [mod] - > ADJ; ADJ[upos = \\ ADJ]; DEM >> N; NUM << N; ADJ >> N; ADJ << DEM\} \end{array} \quad (21.248)$$

(21.248) captures the order of placing first numerals, then nouns, then adjectives, and, finally, demonstratives.

$$\begin{array}{l} \%20 - 3 - 19 - NumDemNounAdj \text{ pattern}\{N[upos = NOUN]; N - \\ [det] - > DEM; DEM[upos = DET, PronType = Dem]; N - [mod] - > \\ NUM; NUM[upos = NUM]; N - [mod] - > ADJ; ADJ[upos = \\ ADJ]; DEM << N; NUM << N; ADJ >> N; NUM << DEM\} \end{array} \quad (21.249)$$

(21.249) captures the order of placing first numerals, then demonstratives, then nouns, and finally, adjectives.

$$\begin{array}{l} \%20 - 3 - 20 - NumAdjNounDem \text{ pattern}\{N[upos = NOUN]; N - \\ [det] - > DEM; DEM[upos = DET, PronType = Dem]; N - [mod] - > \\ NUM; NUM[upos = NUM]; N - [mod] - > ADJ; ADJ[upos = \\ ADJ]; DEM >> N; NUM << N; ADJ << N; NUM << ADJ\} \end{array} \quad (21.250)$$

(21.250) captures the order of placing first numerals, then adjectives, then nouns, and, finally, demonstratives.

$$\begin{array}{l} \%20 - 3 - 21 - AdjNounDemNum \text{ pattern}\{N[upos = NOUN]; N - \\ [det] - > DEM; DEM[upos = DET, PronType = Dem]; N - [mod] - > \\ NUM; NUM[upos = NUM]; N - [mod] - > ADJ; ADJ[upos = \\ ADJ]; DEM >> N; NUM >> N; ADJ << N; DEM << NUM\} \end{array} \quad (21.251)$$

(21.251) captures the order of placing first adjectives, then nouns, then demonstratives, and, finally, numerals.

$$\begin{array}{l} \%20 - 3 - 22 - AdjNounNumDem \text{ pattern}\{N[upos = NOUN]; N - \\ [det] - > DEM; DEM[upos = DET, PronType = Dem]; N - [mod] - > \\ NUM; NUM[upos = NUM]; N - [mod] - > ADJ; ADJ[upos = \\ ADJ]; DEM >> N; NUM >> N; ADJ << N; NUM << DEM\} \end{array} \quad (21.252)$$

(21.252) captures the order of placing first adjectives, then nouns, then numerals, and, finally, demonstratives.

$$\begin{array}{l} \%20 - 3 - 23 - AdjDemNounNum \text{ pattern}\{N[upos = NOUN]; N - \\ [det] - > DEM; DEM[upos = DET, PronType = Dem]; N - [mod] - > \\ NUM; NUM[upos = NUM]; N - [mod] - > ADJ; ADJ[upos = \\ ADJ]; DEM << N; NUM >> N; ADJ << N; ADJ << DEM\} \end{array} \quad (21.253)$$

(21.253) captures the order of placing first adjectives, demonstratives, nouns, and numerals.

$$\begin{array}{l} \%20 - 3 - 24 - AdjNumNounDem \text{ pattern}\{N[upos = NOUN]; N - \\ [det] - > DEM; DEM[upos = DET, PronType = Dem]; N - [mod] - > \\ NUM; NUM[upos = NUM]; N - [mod] - > ADJ; ADJ[upos = \\ ADJ]; DEM >> N; NUM << N; ADJ << N; ADJ << NUM\} \end{array} \quad (21.254)$$

(21.254) captures the order of placing first adjectives, then numerals, then nouns, and, finally, demonstratives.

We must highlight that it is uncommon to find utterances that combine these four elements simultaneously, probably because of the difficulty of processing. Thus, drawing strong tendencies based on a large amount of data is difficult. However, if we check utterances in combinations of 3 elements (instead of 4), we believe the results would be higher, and we could approach the phenomenon more naturally. All the logical possibilities are:

1. DEM-NUM-NOUN
2. NUM-DEM-NOUN
3. ADJ-NUM-NOUN
4. NUM-ADJ-NOUN
5. ADJ-DEM-NOUN
6. DEM-ADJ-NOUN

7. NOUN-DEM-NUM
8. NOUN-NUM-DEM
9. NOUN-ADJ-NUM
10. NOUN-NUM-ADJ
11. NOUN-ADJ-DEM
12. NOUN-DEM-ADJ
13. DEM-NOUN-NUM
14. NUM-DEM-NOUN
15. ADJ-NOUN-NUM
16. NUM-NOUN-ADJ
17. ADJ-NOUN-DEM
18. DEM-NOUN-ADJ

Again, we take advantage of the structures created previously, but in this case, we distribute the appearance of the elements to 3 instead of 4, with all possible orders. That is to say, in each pattern, one dependency will appear less than those previously presented. All these possible orders can be formalized in dependencies as presented in (21.255) - (21.272).

$$\begin{array}{l}
 \%20 - 2 - 1 - DemNumNoun \\
 pattern\{N[upos = NOUN]; N - [det] - > DEM; DEM[upos = \\
 DET, PronType = Dem]; N - [mod] - > NUM; NUM[upos = \\
 NUM]; DEM << N; NUM << N; DEM << NUM\}
 \end{array}
 \tag{21.255}$$

(21.255) captures the order of placing first demonstratives, then numerals, and, finally, nouns.

$$\begin{array}{l}
 \%20 - 2 - 2 - NumDemNoun \\
 pattern\{N[upos = NOUN]; N - [det] - > DEM; DEM[upos = \\
 DET, PronType = Dem]; N - [mod] - > NUM; NUM[upos = \\
 NUM]; DEM << N; NUM << N; NUM << DEM\}
 \end{array}
 \tag{21.256}$$

(21.256) captures the order of placing first numerals, then demonstratives, and, finally, nouns.

$$\begin{array}{l}
\%20 - 2 - 3 - AdjNumNoun \\
pattern\{N[upos = NOUN]; N - [mod]- > NUM; NUM[upos = \\
NUM]; N - [mod]- > ADJ; ADJ[upos = ADJ]; NUM << \\
N; ADJ << N; ADJ << NUM\}
\end{array} \tag{21.257}$$

(21.257) captures the order of placing first adjectives, then numerals, and, finally, nouns.

$$\begin{array}{l}
\%20 - 2 - 4 - NumAdjNoun \\
pattern\{N[upos = NOUN]; N - [mod]- > NUM; NUM[upos = \\
NUM]; N - [mod]- > ADJ; ADJ[upos = ADJ]; NUM << \\
N; ADJ << N; ADJ >> NUM\}
\end{array} \tag{21.258}$$

(21.258) captures the order of placing first numerals, then adjectives, and, finally, nouns.

$$\begin{array}{l}
\%20 - 2 - 5 - AdjDemNoun \\
pattern\{N[upos = NOUN]; N - [det]- > DEM; DEM[upos = \\
DET, PronType = Dem]; N - [mod]- > ADJ; ADJ[upos = \\
ADJ]; DEM << N; ADJ << N; DEM << ADJ\}
\end{array} \tag{21.259}$$

(21.259) captures the order of placing first adjectives, then demonstratives, and, finally, nouns.

$$\begin{array}{l}
\%20 - 2 - 6 - DemAdjNoun \\
pattern\{N[upos = NOUN]; N - [det]- > DEM; DEM[upos = \\
DET, PronType = Dem]; N - [mod]- > ADJ; ADJ[upos = \\
ADJ]; DEM << N; ADJ << N; DEM >> ADJ\}
\end{array} \tag{21.260}$$

(21.260) captures the order of placing first demonstratives, then adjectives, and, finally, nouns.

$$\begin{array}{l}
\%20 - 2 - 7 - NounDemNum \\
pattern\{N[upos = NOUN]; N - [det]- > DEM; DEM[upos = \\
DET, PronType = Dem]; N - [mod]- > NUM; NUM[upos = \\
NUM]; DEM >> N; NUM >> N; DEM << NUM\}
\end{array} \tag{21.261}$$

(21.261) captures the order of placing first nouns, then demonstratives, and, finally, numerals.

$$\begin{array}{l}
\%20 - 2 - 8 - NounNumDem \\
pattern\{N[upos = NOUN]; N - [det] - > DEM; DEM[upos = \\
DET, PronType = Dem]; N - [mod] - > NUM; NUM[upos = \\
NUM]; DEM >> N; NUM >> N; DEM >> NUM\}
\end{array}
\tag{21.262}$$

(21.262) captures the order of placing first nouns, then numerals, and, finally, demonstratives.

$$\begin{array}{l}
\%20 - 2 - 9 - NounAdjNum \\
pattern\{N[upos = NOUN]; N - [mod] - > NUM; NUM[upos = \\
NUM]; N - [mod] - > ADJ; ADJ[upos = ADJ]; NUM >> \\
N; ADJ >> N; ADJ << NUM\}
\end{array}
\tag{21.263}$$

(21.263) captures the order of placing first nouns, then adjectives, and, finally, numerals.

$$\begin{array}{l}
\%20 - 2 - 10 - NounNumAdj \\
pattern\{N[upos = NOUN]; N - [mod] - > NUM; NUM[upos = \\
NUM]; N - [mod] - > ADJ; ADJ[upos = ADJ]; NUM >> \\
N; ADJ >> N; ADJ >> NUM\}
\end{array}
\tag{21.264}$$

(21.264) captures the order of placing first nouns, then numerals, and, finally, adjectives.

$$\begin{array}{l}
\%20 - 2 - 11 - NounAdjDem \\
pattern\{N[upos = NOUN]; N - [det] - > DEM; DEM[upos = \\
DET, PronType = Dem]; N - [mod] - > ADJ; ADJ[upos = \\
ADJ]; DEM >> N; ADJ >> N; DEM >> ADJ\}
\end{array}
\tag{21.265}$$

(21.265) captures the order of placing first nouns, then adjectives, and, finally, demonstratives.

$$\begin{array}{l}
\%20 - 2 - 12 - NounDemAdj \\
pattern\{N[upos = NOUN]; N - [det] - > DEM; DEM[upos = \\
DET, PronType = Dem]; N - [mod] - > ADJ; ADJ[upos = \\
ADJ]; DEM >> N; ADJ >> N; DEM << ADJ\}
\end{array}
\tag{21.266}$$

(21.266) captures the order of placing first nouns, demonstratives, and adjectives.

$$\begin{array}{l}
\%20 - 2 - 13 - \textit{DemNounNum} \\
\textit{pattern}\{N[\textit{upos} = \textit{NOUN}]; N - [\textit{det}] - > \textit{DEM}; \textit{DEM}[\textit{upos} = \\
\textit{DET}, \textit{PronType} = \textit{Dem}]; N - [\textit{mod}] - > \textit{NUM}; \textit{NUM}[\textit{upos} = \\
\textit{NUM}]; \textit{DEM} << N; \textit{NUM} >> N\}
\end{array}
\tag{21.267}$$

(21.267) captures the order of placing first demonstratives, then nouns, and, finally, numerals.

$$\begin{array}{l}
\%20 - 2 - 14 - \textit{NumDemNoun} \\
\textit{pattern}\{N[\textit{upos} = \textit{NOUN}]; N - [\textit{det}] - > \textit{DEM}; \textit{DEM}[\textit{upos} = \\
\textit{DET}, \textit{PronType} = \textit{Dem}]; N - [\textit{mod}] - > \textit{NUM}; \textit{NUM}[\textit{upos} = \\
\textit{NUM}]; \textit{DEM} >> N; \textit{NUM} << N\}
\end{array}
\tag{21.268}$$

(21.268) captures the order of placing first numerals, then demonstratives, and, finally, nouns.

$$\begin{array}{l}
\%20 - 2 - 15 - \textit{AdjNounNum} \\
\textit{pattern}\{N[\textit{upos} = \textit{NOUN}]; N - [\textit{mod}] - > \textit{NUM}; \textit{NUM}[\textit{upos} = \\
\textit{NUM}]; N - [\textit{mod}] - > \textit{ADJ}; \textit{ADJ}[\textit{upos} = \textit{ADJ}]; \textit{NUM} >> \\
N; \textit{ADJ} << N\}
\end{array}
\tag{21.269}$$

(21.269) captures the order of placing first adjectives, then nouns, and, finally, numerals.

$$\begin{array}{l}
\%20 - 2 - 16 - \textit{NumNounAdj} \\
\textit{pattern}\{N[\textit{upos} = \textit{NOUN}]; N - [\textit{mod}] - > \textit{NUM}; \textit{NUM}[\textit{upos} = \\
\textit{NUM}]; N - [\textit{mod}] - > \textit{ADJ}; \textit{ADJ}[\textit{upos} = \textit{ADJ}]; \textit{NUM} << \\
N; \textit{ADJ} >> N\}
\end{array}
\tag{21.270}$$

(21.270) captures the order of placing first numerals, then nouns, and, finally, adjectives.

$$\begin{array}{l}
\%20 - 2 - 17 - \textit{AdjNounDem} \\
\textit{pattern}\{N[\textit{upos} = \textit{NOUN}]; N - [\textit{det}] - > \textit{DEM}; \textit{DEM}[\textit{upos} = \\
\textit{DET}, \textit{PronType} = \textit{Dem}]; N - [\textit{mod}] - > \textit{ADJ}; \textit{ADJ}[\textit{upos} = \\
\textit{ADJ}]; \textit{DEM} >> N; \textit{ADJ} << N\}
\end{array}
\tag{21.271}$$

(21.271) captures the order of placing first adjectives, then nouns, and, finally, demonstratives.

$$\begin{array}{l}
\%20 - 2 - 18 - DemNounAdj \\
pattern\{N[upos = NOUN]; N - [det]- > DEM; DEM[upos = \\
DET, PronType = Dem]; N - [mod]- > ADJ; ADJ[upos = \\
ADJ]; DEM << N; ADJ >> N\}
\end{array} \tag{21.272}$$

(21.272) captures the order of placing first demonstratives, then nouns, and, finally, adjectives.

21.20.2 Trichotomous Expression Based on Grew-Match Data

Although formalizing this Universal is feasible and possible, as shown in the previous chapter, obtaining occurrences that allow an accurate analysis is much more complicated. This Universal could be a clear example of such claims that are much more feasible to characterize through grammars or questionnaires. It is easy to obtain an invented example of an occurrence with a structure containing the four elements cited in this Universal to determine its direction. On the other hand, it is difficult to find enough occurrences in real texts of this type of case. Given the greater difficulty of processing, it is rare to naturally find constructions containing these three dependents simultaneously with the nominal head (as could be “these three red cars”).

Therefore, since the data obtained from examples were not sufficiently high in the languages, we have chosen to carry out separate checks. From the demonstrative-numeral combination, we do not obtain enough data to make a claim, which does not allow us to prove the Universal proposed by Greenberg literally. From the numeral-adjective combination, we obtain a 77% degree of compliance. From the demonstrative-adjective combination, we obtain a 95% compliance rate.

It is important to note that although we have analyzed the parts of the Universal separately, this does not mean that the check is as Universal 18. The main difference is that in the previous Universal, we checked whether the occurrences of individual dependencies were consistent in their direction. In contrast, in this Universal, we assess the order of dependents in cooccurrence. This causes the results to be different. Such a difference can be highlighted for analysis and comparison.

First, we should note that the cooccurrence of the analyzed items manifests generally lower results than when the items appear separately. This simply indicates that, although Greenberg was right in all his claims when we want to refine the claims further and speak not only of direction but of order within the direction, there are logically a few more counterexamples. Second, the results of numerals and demonstratives are in the same position in both Universals (demonstratives have a higher result). This means that the consistency and relationship between adjectives and

demonstratives seem stronger and closer than the relationship between adjectives and numerals, which exhibits a lower predictive power (although still high, not to be forgotten). This implies a lower relationship between the elements, even though they exist.

We also find, again, different data if we compare the two revised cooccurrences of this Universal and if we calculate it jointly. This is due, as in Universal 18, to the fact that the universe is different, caused by languages that possess two of the elements of one cooccurrence but not some of the elements present in the other, which makes the Universal not applicable in such a language. In this case, we find a 92% compliance of languages containing all the necessary elements. This figure is similar to the 90% obtained in the previous Universal, which allows us to place the Universals at the same level of importance (high).

In a trichotomous expression as in figure 21.50:

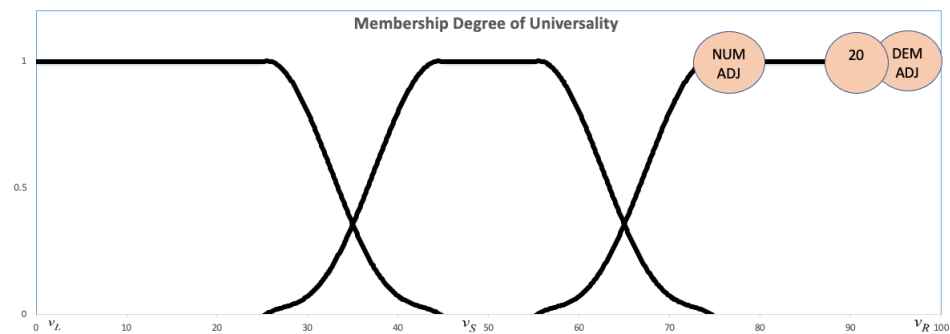


Figure 21.50: Trichotomous Expression for Universal 20

The Universal 20 has been marked in orange as if it were our proposal, as it does not accurately represent the Greenberg claim. Therefore, the three percentages tested are not exact with Greenberg’s claim, but are based on it. In all 3 cases, we find High-Universals with prototypical value at maximum. It seems, therefore, that the orders designated by Greenberg (in cooccurrence, but not with all elements at the same time) follow a fairly systematic order that allows the formulation of rules such as the one reviewed here. We can summarise our findings as in table 21.24.

Summary of characteristics of the Universal 20 analysis:
1. This is an example of a complicated Universal to obtain occurrences in corpus linguistics.
2. The numeral-adjective order has 77% Universality.
3. The demonstrative-adjective order has 95% Universality.
4. The numeral-adjective and demonstrative-adjective orders are High-Universal.

Table 21.24: Summary of Characteristics of the Universal 20 Analysis

21.21 Universal 21

UNIVERSAL 21: *If some or all adverbs follow the adjective they modify, then the language is one in which the qualifying adjective follows the noun, and the verb precedes its nominal object as the dominant order.*

21.21.1 How to Check It in Typometrics

Technical Decisions: 1. Firstly, the Universal is analyzed with Typometrics (Gerdes et al., 2021). 2. We select the SUD annotation scheme because we can get results more comparable to the classical typological claims. 3. We select Graph 2D. 4. We have to generate two different graphs since this Universal is twofold since the condition influences two different linguistic aspects. For the first one, we should select <i>direction – cfc</i> in both x-Axis and y-Axis, meaning that we are interested in the word order feature but applied only in a specific Part of Speech. 5. We set a filter of utterances to avoid misleading data. 6. In the x-Axis we should select <i>ADJ – mod – ADV</i> (since Greenberg explicitly mentions adverbs modifying adjectives) 7. In the y-Axis, we select <i>NOUN – mod – ADJ</i> (since Greenberg explicitly mentions adjectives modifying nouns). 8. For the second one, we select <i>direction – cfc</i> in both, x-Axis and y-Axis too. 9. In the x-Axis, we should select <i>ADJ – mod – ADV</i> again (since it is the shared condition in both cases). 10. In the y-Axis, we select <i>VERB – comp : obj – NOUN</i>, the second correlation (because Greenberg mentions nominal objects).

Table 21.25: Technical Decisions for Checking Universal 21 in Typometrics

21.21.2 Formalization Proposed for Grew-Match

If we want to formally create patterns to represent this Universal (in Grew-Match), we must have six different patterns to contrast (21.273) - (21.278).

$$\begin{array}{l}
\%21 - ADV - bef ADJ \\
pattern\{ADJ[upos = ADJ]; ADJ - [mod]- > ADV; ADV[upos = \\
ADV]; ADJ >> ADV\}
\end{array} \tag{21.273}$$

(21.273) captures adverbs before modifying adjectives.

$$\begin{array}{l}
\%21 - ADV - aft ADJ \\
pattern\{ADJ[upos = ADJ]; ADJ - [mod]- > ADV; ADV[upos = \\
ADV]; ADJ << ADV\}
\end{array} \tag{21.274}$$

(21.274) captures adverbs after modifying adjectives.

$$\begin{array}{l}
\%21 - ADJ - bef N \\
pattern\{N[upos = NOUN]; N - [mod]- > ADJ; ADJ[upos = \\
ADJ]; ADJ << N\}
\end{array} \tag{21.275}$$

(21.275) captures adjectives before modifying nouns.

$$\begin{array}{l}
\%21 - ADJ - aft N \\
pattern\{N[upos = NOUN]; N - [mod]- > ADJ; ADJ[upos = \\
ADJ]; ADJ >> N\}
\end{array} \tag{21.276}$$

(21.276) captures adjectives after modifying nouns.

$$\begin{array}{l}
\%21 - OBJ - bef \\
pattern\{V[upos = VERB]; V - [comp : obj]- > O; O[upos = \\
PROPN|NOUN]; V >> O\} \\
without\{V - [punct]- > P; P[lemma = "?"|"!"]\}
\end{array} \tag{21.277}$$

(21.277) captures objects before verbs.

$$\begin{array}{l}
\%21 - OBJ - aft \\
pattern\{V[upos = VERB]; V - [comp : obj]- > O; O[upos = \\
PROPN|NOUN]; V << O\} \\
without\{V - [punct]- > P; P[lemma = "?"|"!"]\}
\end{array} \tag{21.278}$$

(21.278) captures objects after verbs.

The first two patterns (21.273) and (21.274) focus on the dependency $ADJ - [mod] - > ADV$. The difference between them is the location of the modifying adverb concerning the adjective. In the following two patterns (21.275) and (21.276), a modification is also established, but this time between adjectives and nouns ($N - [mod] - > ADJ$). Again, the difference between the two is the order of the two elements in the sentence. Finally, the last two patterns (21.277) and (21.278) refer to the two possible positions of the direct object in the sentence concerning the verb and, therefore, the proposed dependency is $V - [comp : obj] - > O$, with the two possible orders of $V << O$ and $V >> O$. We should recall that the element V is assigned the value of $[upos = VERB]$ and O of $[upos = NOUN|PROPN]$ to be consistent with the constraints proposed by Greenberg. In addition, we have also added the previously used constraint with the negative pattern *without* to respect the original precept of the dominant (or basic) word order search.

A reasonable doubt a researcher may have is about this last element of the Universal. One possibility is that Greenberg was indeed proposing a Dryer-style Universal, where only the most influential direct opposition in the case of word order (OV/VO) was considered. Another option is that Greenberg refers to the result of this implication being that we are dealing with an SVO, VSO, or VOS language without distinguishing which. Therefore, if one wants to check the different basic word order (and repeat the steps discussed in Universal 1, in section (21.1)) and see if the Universal also applies in these cases, we have to rescue the patterns from Universal 1: (21.279) - (21.284).

%21 – SVO $pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = PROPN NOUN]; V - [comp : obj] - > O; O[upos = PROPN NOUN]; V >> S; S << O; V << O\}$ $without\{V - [punct] - > P; P[lemma = \text{"?" \"!\"}]\}$	(21.279)

(21.279) captures the SVO word order. We ask the subject to be before the verb and the object to be after the verb.

%21 – SOV $pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = PROPN NOUN]; V - [comp : obj] - > O; O[upos = PROPN NOUN]; V >> S; S << O; V >> O\}$ $without\{V - [punct] - > P; P[lemma = \text{"?" \"!\"}]\}$	(21.280)

(21.280) captures the SOV word order. We ask the subject to be before the verb and the object before the verb but not before the subject.

$$\begin{aligned}
 &\%21 - VSO \\
 &pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = \\
 &PROPN|NOUN]; V - [comp : obj] - > O; O[upos = \\
 &PROPN|NOUN]; V << S; S << O\} \\
 &without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
 \end{aligned} \tag{21.281}$$

In the VSO case, in (21.281), we want a verb before the subject and before the object, but the subject before the object.

$$\begin{aligned}
 &\%21 - VOS \\
 &pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = \\
 &PROPN|NOUN]; V - [comp : obj] - > O; O[upos = \\
 &PROPN|NOUN]; V << S; S >> O; V << O\} \\
 &without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
 \end{aligned} \tag{21.282}$$

In VOS word order, in (21.282), we formalize the restriction of verb being before the subject and before the object, but the object appearing before the subject.

$$\begin{aligned}
 &\%21 - OSV \\
 &pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = \\
 &PROPN|NOUN]; V - [comp : obj] - > O; O[upos = \\
 &PROPN|NOUN]; V >> S; S >> O; V >> O\} \\
 &without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
 \end{aligned} \tag{21.283}$$

In the OSV order, in (21.283), we ask for objects appearing before subjects and verbs, but subjects before verbs.

$$\begin{aligned}
 &\%21 - OVS \\
 &pattern\{V[upos = VERB]; V - [subj] - > S; S[upos = \\
 &PROPN|NOUN]; V - [comp : obj] - > O; O[upos = \\
 &PROPN|NOUN]; V << S; S >> O; V >> O\} \\
 &without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
 \end{aligned} \tag{21.284}$$

Moreover, in OVS order, in (21.284), we ask for objects appearing before subjects and verbs, but

verbs before subjects.

Moreover, since this Universal mentions the relation of adjectives to nouns as modifiers, we can also check what happens in situations similar to those developed in Universal 19 (21.19) since it also works with adjectives as modifiers of nouns. Therefore, we decided to reuse (and therefore, it is not necessary to re-explain how it is done) the formalisms (21.285) - (21.296).

$$\begin{array}{l} \%21 - ADJ - aftNUM \\ pattern\{NUM[upos = NUM]; NUM - [mod]- > ADJ; ADJ[upos = \\ ADJ]; ADJ >> NUM\} \end{array} \quad (21.285)$$

(21.285) captures adjectives after the modified numeral.

$$\begin{array}{l} \%21 - ADJ - befNUM \\ pattern\{NUM[upos = NUM]; NUM - [mod]- > ADJ; ADJ[upos = \\ ADJ]; ADJ << NUM\} \end{array} \quad (21.286)$$

(21.286) captures adjectives before the modified numeral.

$$\begin{array}{l} \%21 - ADJ - befPRON \\ pattern\{PRON[upos = PRON]; PRON - [mod]- > \\ ADJ; ADJ[upos = ADJ]; ADJ << PRON\} \end{array} \quad (21.287)$$

(21.287) captures adjectives before the modified pronoun.

$$\begin{array}{l} \%21 - ADJ - aftPRON \\ pattern\{PRON[upos = PRON]; PRON - [mod]- > \\ ADJ; ADJ[upos = ADJ]; ADJ >> PRON\} \end{array} \quad (21.288)$$

(21.288) captures adjectives after the modified pronoun.

$$\begin{array}{l} \%21 - ADJ - aftPROP N \\ pattern\{PROP N[upos = PROP N]; PROP N - [mod]- > \\ ADJ; ADJ[upos = ADJ]; ADJ >> PROP N\} \end{array} \quad (21.289)$$

(21.289) captures adjectives after the modified proper noun.

$$\begin{array}{l}
\%21 - ADJ - bef PROP N \\
pattern\{PROP N[upos = PROP N]; PROP N - [mod] - > \\
ADJ; ADJ[upos = ADJ]; ADJ << PROP N\}
\end{array} \quad (21.290)$$

(21.290) captures adjectives before the modified proper noun.

$$\begin{array}{l}
\%21 - ADJ - bef VERB \\
pattern\{VERB[upos = VERB]; VERB - [comp : pred] - > \\
ADJ; ADJ[upos = ADJ]; ADJ << VERB\}
\end{array} \quad (21.291)$$

(21.291) captures adjectival predicatives before their verb.

$$\begin{array}{l}
\%21 - ADJ - aft VERB \\
pattern\{VERB[upos = VERB]; VERB - [comp : pred] - > \\
ADJ; ADJ[upos = ADJ]; ADJ >> VERB\}
\end{array} \quad (21.292)$$

(21.292) captures adjectival predicatives after their verb.

$$\begin{array}{l}
\%21 - ANY - bef N \\
pattern\{N[upos = NOUN]; N - [mod] - > ANY; \\
ANY[upos = ADJ|ADV|INTJ|NOUN|PROP N|VERB| \\
ADP|AUX|CCONJ|DET|NUM|PART|PRON|SCONJ| \\
SYM]; ANY << N\}
\end{array} \quad (21.293)$$

(21.293) captures any part of speech before the modified noun.

$$\begin{array}{l}
\%21 - ANY - aft N \\
pattern\{N[upos = NOUN]; N - [mod] - > ANY; \\
ANY[upos = ADJ|ADV|INTJ|NOUN|PROP N|VERB| \\
ADP|AUX|CCONJ|DET|NUM|PART|PRON|SCONJ| \\
SYM]; ANY >> N\}
\end{array} \quad (21.294)$$

(21.294) captures any part of speech after the modified noun.

$$\begin{array}{l}
 \%21 - ADJ - bef ANY \\
 pattern\{ADJ[upos = ADJ]; ANY - [mod]- > ADJ; \\
 ANY[upos = ADJ|ADV|INTJ|NOUN|PROPN|VERB| \\
 ADP|AUX|CCONJ|DET|NUM|PART|PRON|SCONJ| \\
 SYM]; ANY >> ADJ\}
 \end{array}
 \tag{21.295}$$

(21.295) captures adjectives before any part of speech is modified.

$$\begin{array}{l}
 \%21 - ADJ - aft ANY \\
 pattern\{ADJ[upos = ADJ]; ANY - [mod]- > ADJ; \\
 ANY[upos = ADJ|ADV|INTJ|NOUN|PROPN|VERB| \\
 ADP|AUX|CCONJ|DET|NUM|PART|PRON|SCONJ| \\
 SYM]; ANY << ADJ\}
 \end{array}
 \tag{21.296}$$

(21.296) captures adjectives after any part of speech is modified.

21.21.3 Graphs from Typometrics

In this Universal, we find again a case of interesting prediction weighed down by a bad formulation. Two requirements must be met for a Universal to be valid in a language. The first of these can be seen in the figure 21.51.

Since the Universal includes the labels “some or all”, we must take into account this Universal language that shows occurrences on the horizontal axis of a postpositional adjective modifying adverbs (from 1 to 100). The limit of the vertical axis should be placed to delimit the category of languages that tend to place the adjective before the noun. Dividing the graph into three parts (noun-modifying adjectives with a tendency to precede, without a dominant order, and with a tendency to place after), we must place ourselves in the lower part, corresponding to the case above.

As can be seen, although there is a clear tendency for the languages to be located in the corner corresponding to 0 values on both axes, many languages fall within the square that should be empty. It is also interesting, however, to note that different languages appear in the upper right part of the graph, while remarkably and clearly, only Sanskrit appears in the designed square. There is a kind of crescent (with a void in the middle and lower right part) which indicates that Greenberg’s proposal is pointing to a rather interesting restriction of languages (although not in a

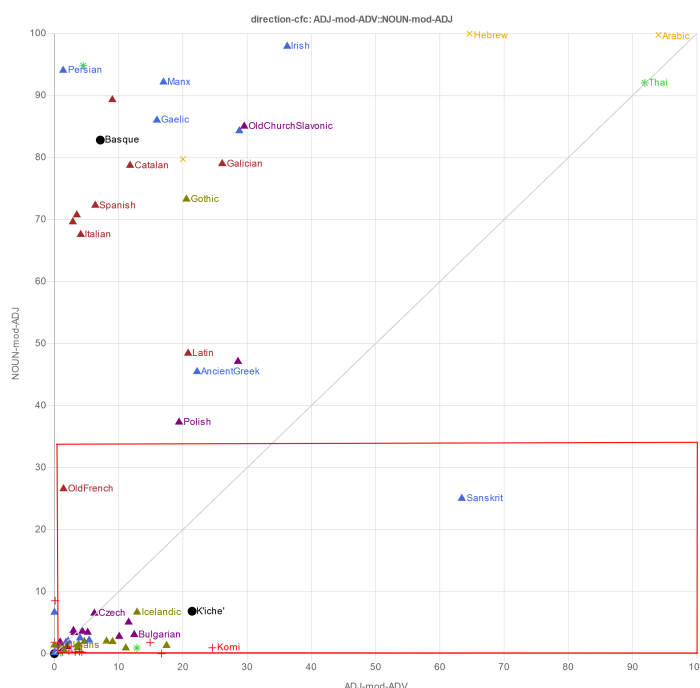


Figure 21.51: Adverbs Modifying Adjectives and Adjectives Modifying Nouns Direction Checking (precise way).

Therefore, we propose a new interpretation of the Universal, which can be seen graphically in the following figure 21.52.

In this case, we propose that the square that is expected to be empty occupies the spaces between that, corresponding to tending to place adverbs behind the adjective they modify and a tendency of NDO on the horizontal axis. On the vertical axis, as in the original Universal, we continue to restrict only languages that place the adjective in front of the noun they modify. The result of such a change of approach means that we would only find Sanskrit as an exception in this first part of the Universal.

The second part of the Universal can be verified using the graph in the figure 21.53.

It can be seen how the predictive power is much more stable in this case. This is mainly due to the rigidity of OV languages in many of their associated parameters, as opposed to the relative flexibility exhibited by VO languages. In this case, if we do not count Sanskrit, which was already part of the exception in the previous part, and Turkish-German because it is Code-Switching, we only find three languages that are an exception to this Universal. In other words, languages that tend to place the nominal object before the verb do not show occurrences of adverbs modifying

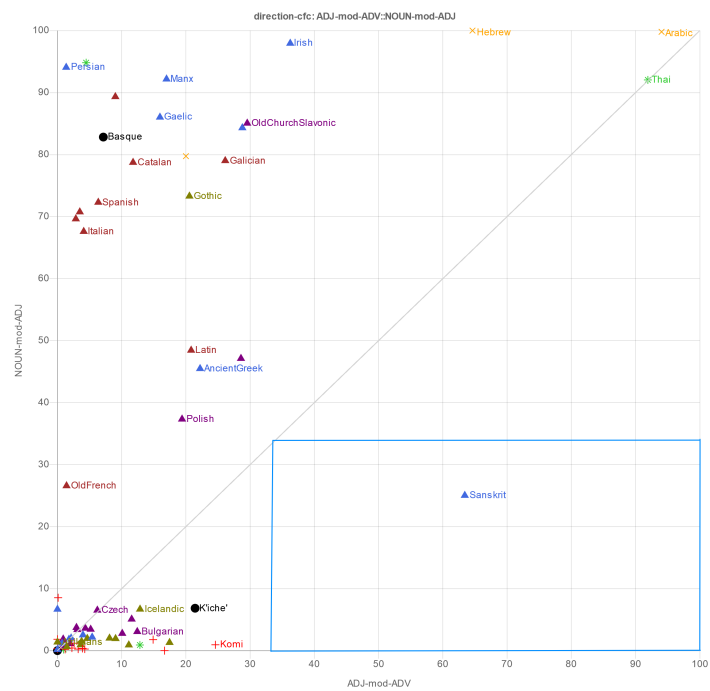


Figure 21.52: Adverbs Modifying Adjectives and Adjectives Modifying Nouns Direction Proposal

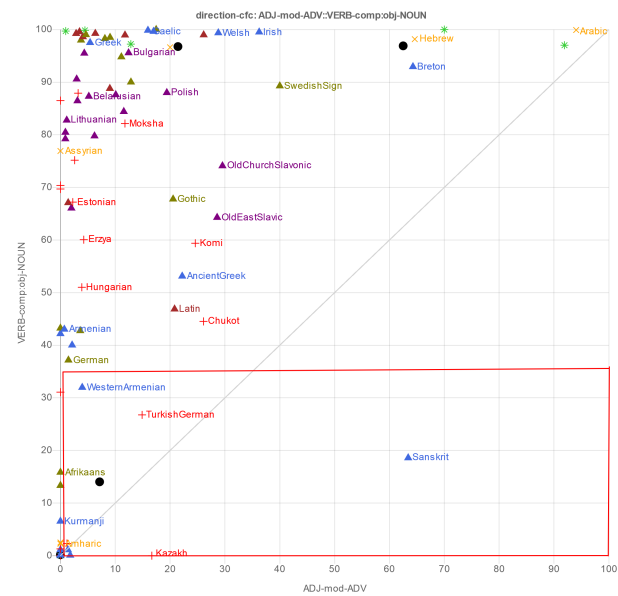


Figure 21.53: Adverbs Modifying Adjectives and Direct Objects Direction Checking

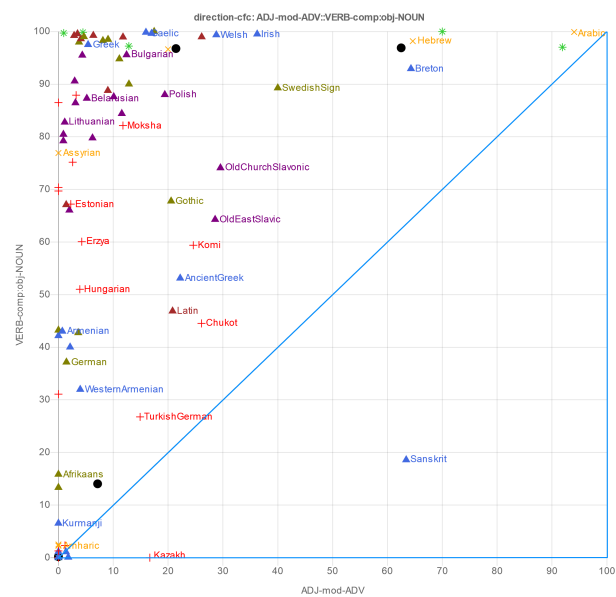


Figure 21.54: Adverbs Modifying Adjectives and Direct Objects Direction Proposal

adjectives behind them.

More clearly, we dare to propose a possible reading in terms of Typometrics about the distribution of the two variables, as shown in figure 21.54.

This proposal shows how all languages (except Kazakh and Sanskrit) are located in the triangle at the upper left. Therefore, a new High-Universal can be asserted in quantitative or Typometrics terms: the world’s languages always show a higher proportion of nominal direct objects after the head than adjectival modifier adverbs.

21.21.4 Trichotomous Expression Based on Grew-Match Data

Through the Grew-Match analysis, we can obtain the following:

- 1. The complete results of the fulfillment of the Universal, which stands at 32%.
- 2. This result is the result of the fulfillment of the first part of the Universal (adverbs following adjectives and adjectives following nouns), which is fulfilled in 35% of the cases, and the second part (adverbs following adjectives and verbs preceding objects) in 73% of the cases.
- 3. We can also check what the results would be if the restriction of “nominal object” were changed to simply “object”, taking into account any part of speech (63% compliance) and if

the Universal in general required the linguistic type SVO instead of VO (going from 32% to 26%).

These data, translated to the trichotomous expression, are shown in figure 21.55.

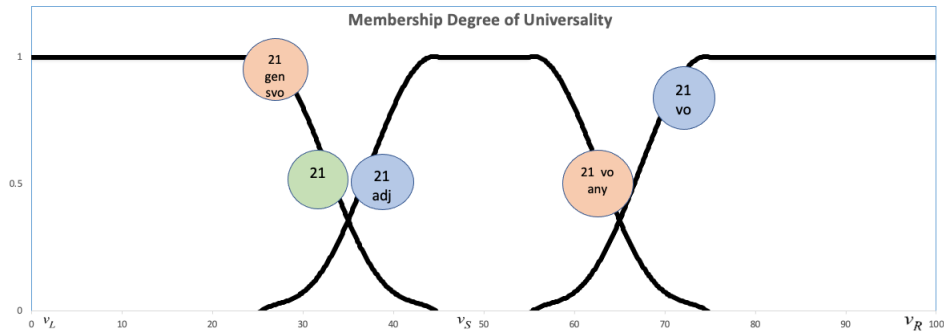


Figure 21.55: Trichotomous Expression for Universal 21

Once the model has been consulted, multiple conclusions can be drawn. First, the complete Universal formulated by Greenberg is in the zone between Low-Universal and Mid-Universal, which means that it cannot be considered to explain the functioning of the world's languages. On the other hand, we see that the reason why this Universal fails in its predictive purpose is the first part, referring to the direction of adjectives modifying nouns, which is also located near that intersection but closer to being considered a Mid-Universal with a tendency to the low part. In contrast, the second part of the Universal is located in the High-Universal part and close to the maximum prototypicality value of the group. Therefore, we propose a modification of the Universal formulated by Greenberg. The first possibility could be to eliminate the first part of the Universal and remain only with this part marked in blue, “21 vo”, corresponding to the second sub-Universal. The second possibility could be to carry out the modification formulated above, in which we stop talking about “some or all” and start considering only languages whose tendency about the dependency of adjectives modified by adjectives is to place the modifiers after the head or which do not show a dominant order. In such a case, the result would be precisely the same as shown in “21 vo” since Sanskrit's only exception would already be included in such a previous Universal.

In addition, we also find it interesting to see how checking modifications of the second part of the Universal shows less convenient results. Both the extension to VO, whose direct object is any part of speech and the change of revision from VO to SVO show how Greenberg's concrete formulation, in this case, is the best possible and the most reliable. We can summarise our findings

as in table 21.26.

Summary of characteristics of the Universal 21 analysis: 1. Typometrics: Except in Sanskrit, no flexibility is observed in the order of the adverb as a modifier of the adjective in languages with the modifier adjective preceding the noun. 2. Typometrics: Except in Sanskrit, no flexibility is observed in the order of the adverb as a modifier of the adjective in OV-type languages. 3. Typometrics: All languages except Kazakh and Sanskrit have more post-verbal nominal direct objects than post-verbal adverbs. 4. Greenberg’s Universal is met 32% of the time. 5. The result of the fulfillment of the Universal is conditioned by the 35% fulfillment of the Universal in part relating adverbs following adjectives to adjectives following nouns. On the other hand, the part of adverbs following adjectives related to verbs preceding objects has a 73% compliance. 6. If we remove the restriction of a part of speech in direct complement nominals, we obtain a compliance of 63%. If we talk about the linguistic type SVO instead of the direction of the object, we obtain a compliance of 26%. 7. The Universal formulated by Greenberg is a low-Universal. The SVO type check, too. 8. The part concerning the direction of adjectives and the proven modification of revising any part of speech are mid-Universals. 9. The part of Greenberg’s Universal concerning the direction of the object and its linkage with the direction of the adverb is a High-Universal.

Table 21.26: Summary of Characteristics of the Universal 21 Analysis

21.22 Universal 22

UNIVERSAL 22: *If in comparisons of superiority, the only order, or one of the alternative orders, is standard-marker-adjective, then the language is postpositional. With overwhelmingly more than chance frequency, the language is prepositional if the only order is adjective-marker-standard.*

21.22.1 Formalization Proposed for Grew-Match

This implicational Universal relates exhibiting a specific structure in superiority comparisons to having prepositions or postpositions. As is well known, knowing the tendency to locate adpositions in a language does not offer difficulties in UD. However, the other part of the Universal is very complicated to obtain. We believe that we cannot offer data exactly faithful to Greenberg’s proposal but rather an approximation that, in our opinion, is still as relevant. Moreover, we can go even further and try to see which possible combinations are reflected in the world’s languages and which seem unproductive without further implications.

First, we must emphasize that UD does not offer the possibility of isolating comparisons of superiority from those of inferiority or equality. However, there is a differentiation between the superiority/inferiority and equality groups, the first group being grouped under “Degree=Cmp” and the second under “Degree=Equ”. In principle, we understand there should be little syntactic difference between superiority and inferiority. However, we should note that the results discussed here will unify occurrences of both types at the same time. We believe this is positive since perhaps we can extend the Universal to cases of inferiority. However, we are not checking this Universal as faithfully as possible due to the limitations of UD. We will offer the same formalizations for the equality cases. Thus, we can see if there are differences between the two groups, and an extension is also possible.

Secondly, we should mention the difficulties in finding a precise formalization in Greenberg’s terms that Universally applies to the world’s languages. He distinguishes three basic elements in superiority comparisons (actually, of any kind): the standard of the comparison, the marker of the superiority comparison, and the adjective referred to.

In reference to the adjective, we have no equivalence problem since the UD annotation scheme contains the same label and applies naturally to all the languages analyzed as a comparative

concept (although its exact realization differs slightly).

About the marker, we find an easily applicable part and another part with limits specific to the tool we use, which we have already discussed previously. Again, different languages can use different strategies to mark a superiority comparison. To the best of our knowledge, we can summarize it in two main groups: languages that use a word to mark it and languages that use a morpheme. We have found that this fact is assigned to adverbs when using a word. Therefore, we propose MARK [upos=ADV, Degree=Cmp] as representative of this option. In the case of morpheme use, we have not been able to check this in other Universals because of the abandonment of pure morphological knowledge in UD. However, in this case, we have, precisely, the *Degree = Cmp* tag, which we can apply to any other element, and it will show the degree of this word (even if we do not know the morpheme that is causing this degree change). Therefore, in the case of these languages, we can propose not to search for a marker isolated from the adjective but a search for an adjective merged with a marker, as in ADJMARK [upos=ADJ, Degree=Cmp]. An example of the first case would be the English construction “more important than”, while the second case can also be found in English in “healthier than”. These cases in which the marker is included in the adjective have not been contemplated by Greenberg since he only conceived a structure of three separate elements. Therefore, these data can be treated separately or together to analyze aspects not included in the Universal (which we believe to be relevant).

The last element, also controversial, is the standard of comparison. In this case, we have the problem of assigning the specific part of speech to such an element, equivalent to the English ‘than.’ The unequivocal case is when experts label it as SCONJ, as may be the case of treebanks in Spanish with the “que”. More problematic is when it is sometimes assigned an ADP label, such as English ‘than.’ In these cases, there may be instances where we are not technically talking about a comparison, such as an example we found “more relevant to female audiences”. However, although we must warn of these influences, we believe that it is still equally interesting to perform this search since we believe that the behavior of the adposition, whether as a standard of comparison or for these cases that are technically not a comparison of superiority, is the same and, therefore, the trend is not blurred. Unfortunately, we refuse to propose an investigation similar to the one proposed in Universal 14, with a manual check of the lemmas, because depending on the language, we obtain different lemmas for the standard of comparison depending on the type of structure or comparison, which would make the work more difficult.

We especially believe that we obtain significant results if we propose a search not restricted by order but by adjacency. Although we will propose both patterns to show how a search could be

done with these, we believe that adjacency, in this case, is more beneficial. Let us give freedom in this sense to the comparison components, having extended the part of speech from standard to adposition. We will find many casual occurrences in which an adposition appears after an adjective. On the other hand, if we demand that all these structures appear together in their entirety, all the cases we find will respond to the equivalent of the comparative structure of superiority or similar structures.

However, we should mention that all we are trying to do is to make a quantitative check of Greenberg's proposal. In no case do we validate the comparison structure he proposed as Universal, not only because many languages merge the adjective with the marker but also because some languages may have other superior comparison strategies different from those contemplated by Greenberg. A clear example would be some Austronesian languages, such as Warlpiri, in which conjoined comparatives can be expressed. The way of expressing that "this dog is bigger than that dog", literally, in English, would be: "This dog is big, this dog is small". In this case, although it is possible to force an interpretation of the categories proposed by Greenberg to fit, this would not be very natural and real since, as we see, Greenberg did not have this type of structure in mind. In short, we believe that the scope of Greenberg's proposal is limited and that some languages may not exhibit any results. However, we should not believe that whenever we do not find occurrences, it is due to the use of a different structure since, in some cases, the absence of occurrences may be due to the type of corpus or the small size of the corpus.

Furthermore, we should note that the implication with postpositions is based on the premise that the stated order is the only or one of the alternative ones (not the dominant or more typical). Regarding implication with prepositions, it is formulated regarding order exclusivity. However, its application does not seem to be total but statistical since Greenberg contemplates some exceptions (without indicating their exact degree, as we already know).

Regarding the order of the adposition, we can reuse patterns already used in Universal 3. Even we can check alternative ways of checking adpositions as already stated. That is, we can see patterns (21.49) - (21.56).

Therefore, we first present the formalizations that would correspond to the different orders if the language makes use of SCONJ for the standard and if it has an isolated word as a marker in an adjacent structure, (21.297) - (21.302).

$$\begin{array}{l}
\%22 - MARK - ADJ - STAND \\
pattern\{MARK[upos = ADV, Degree = Cmp]; ADJ[upos = \\
ADJ]; STAND[upos = CONJ]; MARK < ADJ; ADJ < STAND\}
\end{array} \tag{21.297}$$

(21.297) captures the order of Marker + Adjective + Standard of comparison.

$$\begin{array}{l}
\%22 - ADJ - MARK - STAND \\
pattern\{MARK[upos = ADV, Degree = Cmp]; ADJ[upos = \\
ADJ]; STAND[upos = CONJ]; ADJ < MARK; MARK < \\
STAND\}
\end{array} \tag{21.298}$$

(21.298) captures the order of Adjective + Marker + Standard of comparison.

$$\begin{array}{l}
\%22 - ADJ - STAND - MARK \\
pattern\{MARK[upos = ADV, Degree = Cmp]; ADJ[upos = \\
ADJ]; STAND[upos = CONJ]; ADJ < STAND; STAND < \\
MARK\}
\end{array} \tag{21.299}$$

(21.299) captures the order of Adjective + Standard of comparison + Marker.

$$\begin{array}{l}
\%22 - MARK - STAND - ADJ \\
pattern\{MARK[upos = ADV, Degree = Cmp]; ADJ[upos = \\
ADJ]; STAND[upos = CONJ]; MARK < STAND; STAND < \\
ADJ\}
\end{array} \tag{21.300}$$

(21.300) captures the order of Marker + Standard of comparison + Adjective.

$$\begin{array}{l}
\%22 - STAND - MARK - ADJ \\
pattern\{MARK[upos = ADV, Degree = Cmp]; ADJ[upos = \\
ADJ]; STAND[upos = CONJ]; STAND < MARK; MARK < \\
ADJ\}
\end{array} \tag{21.301}$$

(21.301) captures the order of Standard of comparison + Marker + Adjective.

$$\begin{array}{l}
\%22 - STAND - ADJ - MARK \\
pattern\{MARK[upos = ADV, Degree = Cmp]; ADJ[upos = \\
ADJ]; STAND[upos = CONJ]; STAND < ADJ; ADJ < MARK\}
\end{array} \tag{21.302}$$

(21.302) captures the order of Standard of comparison + Adjective + Marker.

If we want to check the same mentioned in previous patterns but in languages that have the marker inside the adjective, we propose these two possibilities (21.303) and (21.304).

$$\begin{array}{l} \%22 - ADJMARK - STAND \\ pattern\{ADJMARK[upos = ADJ, Degree = Cmp]; STAND[upos = \\ SCONEJ]; ADJMARK < STAND\} \end{array} \quad (21.303)$$

(21.303) captures the order of Marked adjective + Standard of comparison.

$$\begin{array}{l} \%22 - STAND - ADJMARK \\ pattern\{ADJMARK[upos = ADJ, Degree = Cmp]; STAND[upos = \\ SCONEJ]; ADJMARK > STAND\} \end{array} \quad (21.304)$$

(21.304) captures the order of Standard of comparison + Marked adjective.

As mentioned, we can also propose a review of the equality comparison cases. In this case, it will be enough to change the value “Degree=Cmp” to “Degree=Equ” in the formalizations (21.297) - (21.302). Therefore, we do not consider it necessary to repeat them. We can also change the same in the case of languages with comparisons through 2 words instead of 3. That is, we propose to replace in the (21.303) and (21.304) patterns “Degree=Cmp” by “Degree=Equ” again.

Subsequently, as we have explained, we must check the cases in which we add the possibility that the element that serves as a nexus (or standard of comparison in Greenberg’s terms) may also be an adposition. This pattern allows us to capture the cases of superiority comparison in many other languages, possibly finding some synonymous structure that uses another adposition. For this purpose, we propose patterns (21.305) - (21.310).

$$\begin{array}{l} \%22 - MARK - ADJ - STAND - adp \\ pattern\{MARK[upos = ADV, Degree = Cmp]; ADJ[upos = \\ ADJ]; STAND[upos = SCONEJ|ADP]; MARK < ADJ; ADJ < \\ STAND\} \end{array} \quad (21.305)$$

(21.305) captures the order of Marker + Adjective + Standard of comparison (taking into account adpositions).

$$\begin{array}{l}
\%22 - ADJ - MARK - STAND - adp \\
pattern\{MARK[upos = ADV, Degree = Cmp]; ADJ[upos = \\
ADJ]; STAND[upos = CONJ|ADP]; ADJ < MARK; MARK < \\
STAND\}
\end{array}
\tag{21.306}$$

(21.306) captures the order of Adjective + Marker + Standard of comparison (taking into account adpositions).

$$\begin{array}{l}
\%22 - ADJ - STAND - MARK - adp \\
pattern\{MARK[upos = ADV, Degree = Cmp]; ADJ[upos = \\
ADJ]; STAND[upos = CONJ|ADP]; ADJ < STAND; STAND < \\
MARK\}
\end{array}
\tag{21.307}$$

(21.307) captures the order of Adjective + Standard of comparison (taking into account adpositions) + Marker.

$$\begin{array}{l}
\%22 - MARK - STAND - ADJ - adp \\
pattern\{MARK[upos = ADV, Degree = Cmp]; ADJ[upos = \\
ADJ]; STAND[upos = CONJ|ADP]; MARK < \\
STAND; STAND < ADJ\}
\end{array}
\tag{21.308}$$

(21.308) captures the order of Marker + Standard of comparison (taking into account adpositions) + Adjective.

$$\begin{array}{l}
\%22 - STAND - MARK - ADJ - adp \\
pattern\{MARK[upos = ADV, Degree = Cmp]; ADJ[upos = \\
ADJ]; STAND[upos = CONJ|ADP]; STAND < \\
MARK; MARK < ADJ\}
\end{array}
\tag{21.309}$$

(21.309) captures the order of Standard of comparison (considering adpositions) + Marker + Adjective.

$$\begin{array}{l}
\%22 - STAND - ADJ - MARK - adp \\
pattern\{MARK[upos = ADV, Degree = Cmp]; ADJ[upos = \\
ADJ]; STAND[upos = CONJ|ADP]; STAND < ADJ; ADJ < \\
MARK\}
\end{array}
\tag{21.310}$$

(21.310) captures the order of Standard of comparison (considering adpositions) + Adjective + Marker.

Again, we want to check what happens in languages with two-word superiority comparisons using (21.311) and (21.312).

$$\begin{array}{l} \%22 - ADJMARK - STAND - adp \\ pattern\{ADJMARK[upos = ADJ, Degree = Cmp]; STAND[upos = \\ SCONJ|ADP]; ADJMARK < STAND\} \end{array} \quad (21.311)$$

(21.311) captures the order of Marked adjective + Standard of comparison (taking into account adpositions).

$$\begin{array}{l} \%22 - STAND - ADJMARK - adp \\ pattern\{ADJMARK[upos = ADJ, Degree = Cmp]; STAND[upos = \\ SCONJ|ADP]; ADJMARK > STAND\} \end{array} \quad (21.312)$$

(21.312) captures the order of Standard of comparison (considering adpositions) + Marked adjective.

Again, we propose changing the value “Degree=Cmp” to “Degree=Equ” to obtain results about equality comparisons.

In short, Greenberg’s Universal 22 proposal cannot be analyzed broadly across the world’s languages. The approach responds to a very specific structure of specific languages.

21.22.2 Limitations Based on Grew-Match Data

As we have seen in its formalization, this Universal entails some difficulties. However, the limitations of Universal Dependencies also appear in the results shown. The number of languages is so low and insignificant that it does not allow the presentation of reliable or representative results. The main problem is the lack of labeling of the various comparative structures. This additional information is not marked in many languages, so it does not allow the correct identification of the comparative elements in the languages.

This type of inconsistency results from labeling errors; we believe that they are easily solvable in the future. Therefore, we leave the verification of this Universal for the future once more languages are available and these elements have been correctly marked.

21.23 Universal 23

UNIVERSAL 23: *If in apposition, the proper noun usually precedes the common noun, then the language is one in which the governing noun precedes its dependent genitive. With much better than chance frequency, if the common noun usually precedes the proper noun, the dependent genitive precedes its governing noun.*

21.23.1 How to Partially Check it in Grew-Match

In this Universal, we need to contrast different elements via Grew-Match. The location in this section is justified in that it again requires knowledge of the order of the genitive, something that we have already mentioned requires different considerations, something that forces us to use different methods and tools or even to only partially check its manifestation. In addition, we should mention that this Universal has been quite controversial and polemic. For example, Cinque (2013a) devotes this paper to showing multiple inconsistencies in Greenberg’s proposal, giving examples of different languages whose behavior is the opposite, even if they are languages known by Greenberg himself. In this thesis, we propose to make this critical approach of Cinque more systematic, not only focusing on disordered examples of different languages but also through computational data and real texts in multiple languages. We would like to point out that capturing a canonical order of the common and proper nouns in adpositions is difficult.

The first term to disambiguate is “apposition”. This concept can encompass different cases, which can be misleading and confusing, such as syntagms like “President Biden” or “Frank, the dentist”, In the case of this Universal. However, the formulation is ambiguous; if we read the previous and subsequent development of the Universal in Greenberg’s paper and analyze its examples, we realize that it refers only to the first case, with examples such as “Mr. X” or “Avenue X”. Even in the English language, we can find cases in both senses, for in President Biden, we find the proper noun in the second position, while in Camden Road, the proper noun is in the first. Therefore, we anticipate the results may indicate something different from what Greenberg proposed.

In order to formalize the different options for testing, we must consider the variables to be collected in relation to the appositions. First of all, we must remember that the cases Greenberg checked are always a union of a proper noun and a common noun. We must check the situations in which the common or proper name acts as the relevant head or word in order to see what behavior

there is in the languages of the world and to avoid possible errors in the annotator's criteria since in the examples previously discussed; we understand that in "Mr. X", for example, there may be a difference of criteria in the annotation. Secondly, we must consider whether it is more interesting to formalize it with direction and adjacency (" $<$ " and " $>$ ") or simply with direction (" $>>$ " and " $<<$ "). Once we carry out the investigation, we will see the behavior of these cases and if the restriction to proximity as it seems to exist gives more fruitful results or, on the contrary, there are cases in which there are elements between the two nouns and, therefore, it is better to speak of direction in general. Finally, we must collect both directions (the head to the left or right). In short, the patterns we propose for this check are (21.313) - (21.320).

$$\begin{array}{l} \%23 - APPOS - PNh - befCN - ad \\ pattern\{CN[upos = NOUN]; PN[upos = PROPN]; PN - [appos] - > \\ CN; PN < CN\} \end{array} \quad (21.313)$$

(21.313) captures appositions of depending common nouns in adjacency after proper nouns.

$$\begin{array}{l} \%23 - APPOS - PNh - aftCN - ad \\ pattern\{CN[upos = NOUN]; PN[upos = PROPN]; PN - [appos] - > \\ CN; PN > CN\} \end{array} \quad (21.314)$$

(21.314) captures appositions of depending common nouns in adjacency before proper nouns.

$$\begin{array}{l} \%23 - APPOS - PNh - befCN - Nad \\ pattern\{CN[upos = NOUN]; PN[upos = PROPN]; PN - [appos] - > \\ CN; PN << CN\} \end{array} \quad (21.315)$$

(21.315) captures appositions of depending common nouns after proper nouns.

$$\begin{array}{l} \%23 - APPOS - PNh - aftCN - Nad \\ pattern\{CN[upos = NOUN]; PN[upos = PROPN]; PN - [appos] - > \\ CN; PN >> CN\} \end{array} \quad (21.316)$$

(21.316) captures appositions of depending common nouns before proper nouns.

$$\begin{array}{l} \%23 - APPOS - CNh - befPN - ad \\ pattern\{CN[upos = NOUN]; PN[upos = PROPN]; CN - [appos] - > \\ PN; PN > CN\} \end{array} \quad (21.317)$$

(21.317) captures appositions of depending proper nouns in adjacency after common nouns.

$$\begin{array}{l} \%23 - APPOS - CNh - aftPN - ad \\ pattern\{CN[upos = NOUN]; PN[upos = PROPN]; CN - [appos] - > \\ PN; PN < CN\} \end{array} \quad (21.318)$$

(21.318) captures appositions of depending proper nouns in adjacency before common nouns.

$$\begin{array}{l} \%23 - APPOS - CNh - befPN - Nad \\ pattern\{CN[upos = NOUN]; PN[upos = PROPN]; CN - [appos] - > \\ PN; PN >> CN\} \end{array} \quad (21.319)$$

(21.319) captures appositions of depending proper nouns after common nouns.

$$\begin{array}{l} \%23 - APPOS - CNh - aftPN - Nad \\ pattern\{CN[upos = NOUN]; PN[upos = PROPN]; CN - [appos] - > \\ PN; PN << CN\} \end{array} \quad (21.320)$$

(21.320) captures appositions of depending proper nouns before common nouns.

The second part of the Universal is the one that refers to the genitive and the order of the genitive. In order to check it, we must follow the steps already explained in Universal 2 (21.2). For this reason, we refer to the formalizations (21.321) - (21.334). However, again, we will focus on the genitive cases before extending the results.

$$\begin{array}{l} \%23 - Gen - aft \\ pattern\{N[upos = NOUN]; GEN[Case = Gen]; N - [mod] - > \\ GEN; N << GEN\} \end{array} \quad (21.321)$$

(21.321) captures how the dependent with genitive case appears after the nominal head.

$$\begin{array}{l} \%23 - Gen - bef \\ pattern\{N[upos = NOUN]; GEN[Case = Gen]; N - [mod] - > \\ GEN; N >> GEN\} \end{array} \quad (21.322)$$

(21.322) captures how the dependent with genitive case appears before the nominal head.

$$\begin{array}{l} \%23 - Gen - aft - adp \\ pattern\{N[upos = NOUN]; ADP[upos = ADP]; N - [mod] - > \\ ADP; N << ADP\} \end{array} \quad (21.323)$$

(21.323) captures how a modifying genitive by adposition appears after the nominal head.

$$\begin{array}{l} \%23 - Gen - bef - adp \\ pattern\{N[upos = NOUN]; ADP[upos = ADP]; N - [mod] - > \\ ADP; N >> ADP\} \end{array} \quad (21.324)$$

(21.324) captures how a modifying genitive by adposition appears before the nominal head.

$$\begin{array}{l} \%23 - Gen - aft - adp - udep \\ pattern\{N[upos = NOUN]; ADP[upos = ADP]; N - [udep] - > \\ ADP; N << ADP\} \end{array} \quad (21.325)$$

(21.325) captures how a genitive by adposition appears after the nominal head.

$$\begin{array}{l} \%23 - Gen - bef - adp - udep \\ pattern\{N[upos = NOUN]; ADP[upos = ADP]; N - [udep] - > \\ ADP; N >> ADP\} \end{array} \quad (21.326)$$

(21.326) captures how a genitive by adposition appears before the nominal head.

$$\begin{array}{l} \%23 - Gen - aft - adp - any \\ pattern\{N[upos = NOUN]; ADP[upos = ADP]; N - > ADP; N << \\ ADP\} \end{array} \quad (21.327)$$

(21.327) captures how a genitive with any dependency by adposition appears after the nominal head.

$$\begin{array}{l} \%23 - Gen - bef - adp - any \\ pattern\{N[upos = NOUN]; ADP[upos = ADP]; N - > ADP; N >> \\ ADP\} \end{array} \quad (21.328)$$

(21.328) captures how a genitive with any dependency by adposition appears before the nominal head.

$$\begin{array}{l}
\%23 - Gen - nn - mod - n2aft - ady \\
pattern\{N1[upos = NOUN]; N2[upos = NOUN]; N1 - [mod] - > \\
N2; N1 < N2\}
\end{array}
\quad (21.329)$$

(21.329) captures the adjacency of genitive modifiers after the nominal head.

$$\begin{array}{l}
\%23 - Gen - nn - mod - n2bef - ady \\
pattern\{N1[upos = NOUN]; N2[upos = NOUN]; N1 - [mod] - > \\
N2; N1 > N2\}
\end{array}
\quad (21.330)$$

(21.330) captures the adjacency of genitive modifiers before the nominal head.

$$\begin{array}{l}
\%23 - Gen - nn - comp - n2aft - ady \\
pattern\{N1[upos = NOUN]; N2[upos = NOUN]; N1 - [compund] - > \\
N2; N1 < N2\}
\end{array}
\quad (21.331)$$

(21.331) captures the adjacency of genitive compounds after the nominal head.

$$\begin{array}{l}
\%23 - Gen - nn - comp - n2bef - ady \\
pattern\{N1[upos = NOUN]; N2[upos = NOUN]; N1 - [compound] - > \\
N2; N1 > N2\}
\end{array}
\quad (21.332)$$

(21.332) captures the adjacency of genitive compounds before the nominal head.

$$\begin{array}{l}
\%23 - Gen - nn - appos - n2aft - ady \\
pattern\{N1[upos = NOUN]; N2[upos = NOUN]; N1 - [appos] - > \\
N2; N1 < N2\}
\end{array}
\quad (21.333)$$

(21.333) captures the adjacency of genitive appositions after the nominal head.

$$\begin{array}{l}
\%23 - Gen - nn - appos - n2bef - ady \\
pattern\{N1[upos = NOUN]; N2[upos = NOUN]; N1 - [appos] - > \\
N2; N1 > N2\}
\end{array}
\quad (21.334)$$

(21.334) captures the adjacency of genitive appositions before the nominal head.

21.23.2 Trichotomous Expression Based on Grew-Match Data

The results do not seem too high, indicating a poor formulation of the Universal. Since the only result that stands out positively is that of the adpositional structures of proper nouns preceding common nouns and genitives following the governing noun, we believe that, perhaps, Greenberg started from many examples of this type to formulate his Universal, which may have been misleading.

All these data can be summarized in figure 21.56.

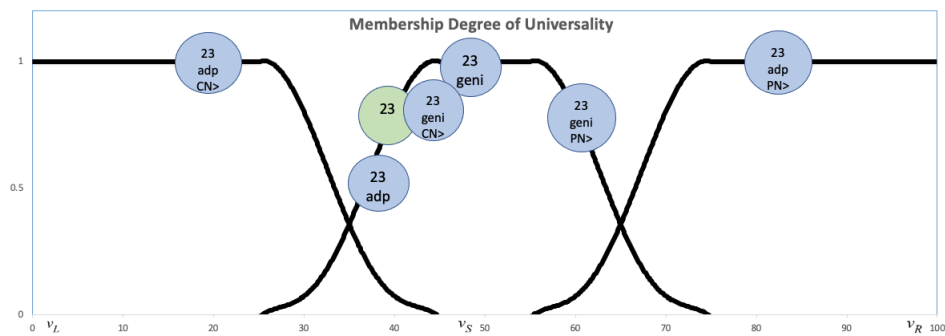


Figure 21.56: Trichotomous Expression for Universal 23

The original Universal and most small parts that form it are placed in the mid-Universal set. This tells us that such a claim cannot be used to understand and predict the behavior of languages in their types. Two parts of the Universal escape the Mid-Universal set. First, formulating adpositional structures with the common noun preceding the appositions is Low-Universal. Second, formulating adpositional structures with the proper noun preceding the appositions is High-Universal. We can summarise our findings as in table 21.27.

Summary of characteristics of the Universal 23 analysis: 1. Compliance with this Universal in its original terms is 42%. 2. We have first broken down the results in reference to languages with genitive case structure and languages with a tendency to greater prepositional use, with results of 47% compliance in the first case and 38% in the second. 3. In addition, we wanted to further detail the fulfillment of this Universal and to see if the two parts of the Universal are equally fulfilled in the two structures announced. About the adpositional structures, we found that the relation between the common noun preceding the proper noun and the genitive preceding the governing noun is fulfilled in 20% of the occasions. 4. The relation between the proper noun preceding the common noun and the genitive following the governing noun is fulfilled in 83% of the cases. 5. For the occurrences with genitive case, we found 42% compliance in the cases of common noun preceding proper noun and 60% in the cases of proper noun preceding common noun. 6. This Universal is of no use and interest for Linguistic Typology as it is formulated in its original terms since it is a Mid-Universal. 7. A reduction of the concrete Universal to the annotated case as the only example of High-Universal is the only possibility to deal with the issue raised by this Universal. 8. The part of the Universal found to be Low-Universal is the main cause of this imbalance and bias in the Universal.

Table 21.27: Summary of Characteristics of the Universal 23 Analysis

21.24 Universal 24

UNIVERSAL 24: *If the relative expression precedes the noun as the only construction or as an alternate construction, either the language is postpositional, or the adjective precedes the noun or both.*

21.24.1 How to Check It in Typometrics

- Technical Decisions:**
1. Firstly, the Universal is analyzed with Typometrics (Gerdes et al., 2021).
 2. We select the SUD annotation scheme because we can get results more comparable to the classical typological claims.
 3. This Universal is complex since it involves different scenarios. For this, we need different 2D graphs.
 4. We set a filter of utterances to avoid misleading data.
 5. For the first graph we use the *direction – cfc* of *NOUN – mod – VERB* in the x-Axis and the *direction – cfc* of *ADP – comp : obj – NOUN* (prepositional or postpositional languages) in the y-Axis.
 6. For the second graph we use the *direction – cfc* of *NOUN – mod – VERB* in the x-Axis again and the *direction – cfc* of *NOUN – mod – ADJ* in the y-Axis. We must consider that a language fits in the Universal if it satisfies 1 of the two designed graphs (or both).

Table 21.28: Technical Decisions for Checking Universal 24 in Typometrics

Therefore, the Universal that we can test in this case using Typometrics is not the one strictly proposed by Greenberg but a much more general kind of adaptation. Some examples of verbs modifying nouns do not necessarily have to correspond to relative clauses and may only be verbs in some languages. Therefore, the use of Typometrics should be done to try to check graphically possible adaptations of this Universal, while we should leave the use of Grew-Match for the faithful representation of the Universal.

In Grew-Match, first of all, we want to strictly reproduce the Greenberg Universal with all the details. To do this, we must capture the position of the relative clause that modifies a noun

concerning the noun.

The first step we had taken was to propose the creation of: $N[upos = NOUN|PROPN]$ and $VERB[upos = VERB|AUX]$. In this case, we added the two noun forms there (common and proper) to the noun tag and added both verb and auxiliary forms to the value of $VERB$ for two reasons. The first one is because the relative clause can be made with what we could call a copulative verb in some languages and also because in SUD, when there are verbs of the type “have eaten”, the dependency of the relative clause falls on the AUX and, instead, the $VERB$ is as a dependent about $comp : aux$. The idea is to ask for a dependency of $N - [mod@relcl] - > VERB$, a formalization corresponding to a noun modified by a relative clause with any type of verb.

One of the dilemmas arises is whether we are incurring in Anglocentrism when considering only those two parts of speech as possible heads of a relative clause. In languages such as Thai, we can find adjectives with verbal characteristics (e.g., having a dependent subject) or, in languages such as Chinese, having a particle as head. Therefore, it would be interesting to consider that the head of the relative clause modifying a noun could be any part of speech, which would correspond to changing the value of the $VERB$ to $ANY[upos = ADJ|ADV|INTJ|NOUN|PROPN|VERB|ADP|AUX|CCONJ|DET|NUM|PART|PRON|SCONJ]$.

Another option we have assessed is that Greenberg’s proposal is too Anglocentric by itself. If we really must work with strictly relative clauses, we need a relative pronoun in such a clause. This would cause us to also have to activate a $REL[upos = Pron, PronType = Rel|“Int, Rel”]$ and also an extra dependency on $ANY - > REL$. Recall that we are not only looking for pronouns with the value Rel but also Int, Rel because there are languages where the differentiation between Int and Rel is not established, and then they share that same label. Our options are to ignore those languages or take that tag into account, with the caveat that the result may be misleading. However, we prefer to include it because we can also automatically discard this double tag and filter it negatively. We always have the option to stay only with the pure Rel . Delete interrogatives by applying the negative filtering *without* that we have shown in previous patterns. We confirm our claims with a specialized bibliography, as can be De Vries (2005, 2018).

All this means there are multiple possibilities to formalize these relative clauses, from the most conservative and close to Greenberg’s Universal to the less Anglocentric but general ones. We must also warn that we have detected a specific error in some treebank (for example, in AnCora, either in Catalan or Spanish) in which the additional information of $@relcl$ is missing in these annotations and, therefore, the dependency is on mod in general, unspecified. The only thing that

can be done in these cases is to change the *mod@relcl* dependency to *mod*, but in order to make sure that the *mod* relation refers to a relative clause, here we are forced to maintain the fact that the *REL* element shown above must be present.

21.24.2 Formalization Proposed for Grew-Match

In summary, the different patterns we could have to capture this Universal (and the different options to solve the problems that some treebanks present in their tagging) are (21.335) - (21.342).

$$\begin{array}{l} \%24 - Rel - Basic - aft \\ pattern\{N[upos = NOUN|PROPN]; N - [mod@relcl]- > \\ VERB; VERB[upos = VERB|AUX]; N << VERB\} \end{array} \quad (21.335)$$

(21.335) captures relative clauses after the modified noun.

$$\begin{array}{l} \%24 - Rel - Basic - bef \\ pattern\{N[upos = NOUN|PROPN]; N - [mod@relcl]- > \\ VERB; VERB[upos = VERB|AUX]; N >> VERB\} \end{array} \quad (21.336)$$

(21.336) captures relative clauses before the modified noun.

$$\begin{array}{l} \%24 - Rel - withPron - aft \\ pattern\{N[upos = NOUN|PROPN]; N - [mod@relcl]- > \\ VERB; VERB[upos = VERB|AUX]; REL[upos = \\ PRON, PronType = Rel|“Int, Rel”]; VERB- > REL; N << VERB\} \\ without\{V - [punct]- > P; P[lemma = “?”|“!”]\} \end{array} \quad (21.337)$$

(21.337) captures relative clauses with explicit pronouns after the modified noun.

$$\begin{array}{l} \%24 - Rel - withPron - bef \\ pattern\{N[upos = NOUN|PROPN]; N - [mod@relcl]- > \\ VERB; VERB[upos = VERB|AUX]; REL[upos = \\ PRON, PronType = Rel|“Int, Rel”]; VERB- > REL; N >> VERB\} \\ without\{V - [punct]- > P; P[lemma = “?”|“!”]\} \end{array} \quad (21.338)$$

(21.338) captures relative clauses with explicit pronouns before the modified noun.

$$\begin{array}{l}
\%24 - Rel - ANY - aft \\
pattern\{N[upos = NOUN|PROPN]; N - \\
[mod@relcl] - > ANY; ANY[upos = \\
ADJ|ADV|INTJ|NOUN|PROPN|VERB|ADP|AUX| \\
CCONJ|DET|NUM|PART|PRON|SCONJ]; N << ANY\}
\end{array} \quad (21.339)$$

(21.339) captures relative clauses with any part of speech as head after the modified noun.

$$\begin{array}{l}
\%24 - Rel - ANY - bef \\
pattern\{N[upos = NOUN|PROPN]; N - \\
[mod@relcl] - > ANY; ANY[upos = \\
ADJ|ADV|INTJ|NOUN|PROPN|VERB|ADP|AUX| \\
CCONJ|DET|NUM|PART|PRON|SCONJ]; N >> ANY\}
\end{array} \quad (21.340)$$

(21.340) captures relative clauses with any part of speech as head before the modified noun.

$$\begin{array}{l}
\%24 - Rel - ANY withPron - aft \\
pattern\{N[upos = NOUN|PROPN]; N - \\
[mod@relcl] - > ANY; ANY[upos = \\
ADJ|ADV|INTJ|NOUN|PROPN|VERB|ADP|AUX| \\
CCONJ|DET|NUM|PART|PRON|SCONJ]; REL[upos = \\
PRON, PronType = Rel["Int, Rel"]; ANY - > REL; N << ANY\} \\
without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
\end{array} \quad (21.341)$$

(21.341) captures relative clauses with any part of speech as head and an explicit pronoun after the modified noun.

$$\begin{array}{l}
\%24 - Rel - ANY withPron - bef \\
pattern\{N[upos = NOUN|PROPN]; N - \\
[mod@relcl] - > ANY; ANY[upos = \\
ADJ|ADV|INTJ|NOUN|PROPN|VERB|ADP|AUX| \\
CCONJ|DET|NUM|PART|PRON|SCONJ]; REL[upos = \\
PRON, PronType = Rel["Int, Rel"]; ANY - > REL; N >> ANY\} \\
without\{V - [punct] - > P; P[lemma = "?"|"!"]\}
\end{array} \quad (21.342)$$

(21.342) captures relative clauses with any part of speech as head and an explicit pronoun before the modified noun.

Without seeing the results we obtain specifically on languages, we cannot decide which is the most appropriate formalization for the problem presented since we do not know whether there may be more annotation errors or other unforeseen complications. Therefore, when presenting the results, we will comment on them.

The next element of the Universal is the analysis of the position of the adposition. As we have already commented above, Greenberg was referring to the relation only to nouns. Therefore, without explaining the design of the patterns, since we have already dealt with them elsewhere in the dissertation in Universal 3 for example, we propose a more faithful approach to Greenberg and a more complete and detailed one, with different parts of speech or with multiple ones, to check if there are also interesting results (21.343) - (21.350).

$$\begin{array}{l} \%24 - Prep - N \\ pattern\{ADP[upos = ADP]; ADP - [comp : obj] - > N; N[upos = \\ NOUN|PROPN]; ADP << N\} \end{array} \quad (21.343)$$

(21.343) captures prepositions + nouns.

$$\begin{array}{l} \%24 - Post - N \\ pattern\{ADP[upos = ADP]; ADP - [comp : obj] - > N; N[upos = \\ NOUN|PROPN]; ADP >> N\} \end{array} \quad (21.344)$$

(21.344) captures nouns + postpositions.

$$\begin{array}{l} \%24 - Prep - PN \\ pattern\{ADP[upos = ADP]; ADP - [comp : obj] - > PN; PN[upos = \\ PRON]; ADP << PN\} \end{array} \quad (21.345)$$

(21.345) captures prepositions + pronouns.

$$\begin{array}{l} \%24 - Post - PN \\ pattern\{ADP[upos = ADP]; ADP - [comp : obj] - > PN; PN[upos = \\ PRON]; ADP >> PN\} \end{array} \quad (21.346)$$

(21.346) captures pronouns + prepositions.

$$\begin{array}{l}
\%24 - \textit{Prep} - V \\
\textit{pattern}\{\textit{ADP}[\textit{upos} = \textit{ADP}]; \textit{ADP} - [\textit{comp} : \textit{obj}] - > V; V[\textit{upos} = \\
\textit{VERB}]; \textit{ADP} << V\}
\end{array} \tag{21.347}$$

(21.347) captures prepositions + depending verbs.

$$\begin{array}{l}
\%24 - \textit{Post} - V \\
\textit{pattern}\{\textit{ADP}[\textit{upos} = \textit{ADP}]; \textit{ADP} - [\textit{comp} : \textit{obj}] - > V; V[\textit{upos} = \\
\textit{VERB}]; \textit{ADP} >> V\}
\end{array} \tag{21.348}$$

(21.348) captures depending verbs + postpositions.

$$\begin{array}{l}
\%24 - \textit{Prep} - \textit{ANY} \\
\textit{pattern}\{\textit{ADP}[\textit{upos} = \textit{ADP}]; \textit{ADP} - \\
[\textit{comp} : \textit{obj}] - > \textit{ANY}; \textit{ANY}[\textit{upos} = \\
\textit{ADJ}|\textit{ADV}|\textit{INTJ}|\textit{NOUN}|\textit{PROPN}|\textit{VERB}|\textit{ADP}|\textit{AUX}| \\
\textit{CCONJ}|\textit{DET}|\textit{NUM}|\textit{PART}|\textit{PRON}|\textit{SCONJ}|\textit{SYM}; \\
\textit{ADP} << \textit{ANY}\}
\end{array} \tag{21.349}$$

(21.349) captures prepositions + any depending part of speech.

$$\begin{array}{l}
\%24 - \textit{Post} - \textit{ANY} \\
\textit{pattern}\{\textit{ADP}[\textit{upos} = \textit{ADP}]; \textit{ADP} - \\
[\textit{comp} : \textit{obj}] - > \textit{ANY}; \textit{ANY}[\textit{upos} = \\
\textit{ADJ}|\textit{ADV}|\textit{INTJ}|\textit{NOUN}|\textit{PROPN}|\textit{VERB}|\textit{ADP}|\textit{AUX}| \\
\textit{CCONJ}|\textit{DET}|\textit{NUM}|\textit{PART}|\textit{PRON}|\textit{SCONJ}|\textit{SYM}; \\
\textit{ADP} >> \textit{ANY}\}
\end{array} \tag{21.350}$$

(21.350) captures any depending part of speech + postpositions.

Finally, we must check the last element of the Universal, related to the position of an adjective modifying a noun. In this case, again, we can take advantage of formalizations already made for other Universals in which this parameter is taken into account (and for this reason, we do not comment on the elaboration process either), regarding Universal 17, for example. The patterns we propose respect Greenberg's formulation and modifications to obtain more complete and representative results for the languages (21.351) - (21.360).

$$\begin{array}{l}
 \%24 - ADJ - after \\
 pattern\{ADJ[upos = ADJ]; N - [mod]- > ADJ; N[upos = \\
 PROP N|NOUN]; N << ADJ\}
 \end{array}
 \quad (21.351)$$

(21.351) captures adjectives after the modified noun.

$$\begin{array}{l}
 \%24 - ADJ - before \\
 pattern\{ADJ[upos = ADJ]; N - [mod]- > ADJ; N[upos = \\
 PROP N|NOUN]; N >> ADJ\}
 \end{array}
 \quad (21.352)$$

(21.352) captures adjectives before the modified noun.

$$\begin{array}{l}
 \%24 - ADJ - PRON - after \\
 pattern\{ADJ[upos = ADJ]; PN - [mod]- > ADJ; PN[upos = \\
 PRON]; PN << ADJ\}
 \end{array}
 \quad (21.353)$$

(21.353) captures adjectives after the modified pronoun.

$$\begin{array}{l}
 \%24 - ADJ - PRON - before \\
 pattern\{ADJ[upos = ADJ]; PN - [mod]- > ADJ; PN[upos = \\
 PRON]; PN >> ADJ\}
 \end{array}
 \quad (21.354)$$

(21.354) captures adjectives before the modified pronoun.

$$\begin{array}{l}
 \%24 - ADJ - predicative - before \\
 pattern\{V[upos = VERB|AUX]; V - [comp : pred]- > \\
 ADJ; ADJ[upos = ADJ]; V >> ADJ\}
 \end{array}
 \quad (21.355)$$

(21.355) captures predicative adjectives before their verb.

$$\begin{array}{l}
 \%24 - ADJ - predicative - after pattern\{V[upos = VERB|AUX]; V - \\
 [comp : pred]- > ADJ; ADJ[upos = ADJ]; V << ADJ\}
 \end{array}
 \quad (21.356)$$

(21.356) captures predicative adjectives after their verb.

%24 – *ADJ* – *any* – *after*

pattern{*ANY*[*upos* = *ADJ*]|*ADV*|*INTJ*|*NOUN*|*PROPN*|

VERB|*ADP*|*AUX*|*CCONJ*|*DET*|*NUM*|*PART*|*PRON*|

SCONJ|*SYM*]; *ANY* – [*mod*] – > *ADJ*; *ADJ*[*upos* = *ADJ*]; *ANY* <<

ADJ}

(21.357)

(21.357) captures adjectives after any part of modified speech.

%24 – *ADJ* – *any* – *before*

pattern{*ANY*[*upos* = *ADJ*]|*ADV*|*INTJ*|*NOUN*|*PROPN*|

VERB|*ADP*|*AUX*|*CCONJ*|*DET*|*NUM*|*PART*|*PRON*|

SCONJ|*SYM*]; *ANY* – [*mod*] – > *ADJ*; *ADJ*[*upos* = *ADJ*]; *ANY* >>

ADJ}

(21.358)

(21.358) captures adjectives before any part of modified speech.

%24 – *ADV* – *after*

pattern{*ADJ*[*upos* = *ADJ*]; *ADJ* – [*mod*] – > *ADV*; *ADV*[*upos* =

ADV]; *ADJ* << *ADV*}

(21.359)

(21.359) captures adverbs after the modified adjective.

%24 – *ADV* – *before*

pattern{*ADJ*[*upos* = *ADJ*]; *ADJ* – [*mod*] – > *ADV*; *ADV*[*upos* =

ADV]; *ADJ* >> *ADV*}

(21.360)

(21.360) captures adverbs before the modified adjective.



Figure 21.57: Verbs Modifying Nouns and Adposition Direction Checking

21.24.3 Graphs from Typometrics

This Universal also requires that we display the data in two independent graphs. In addition, the limitations of Typometrics (not allowing filtering by “mod@relcl” but by “mod”) mean that the results need to be further verified with Grew-Match, which does allow such precision. In this case, some verb constructions may appear as modifiers of a noun that are not relative. In any case, the first part of the Universal is represented as in the figure 21.57.

The fact that the linguistic type of implication is mentioned as a single or alternative order means that we have to consider any language that is not at 100 (absence of a single or alternative order preceded) on the horizontal axis. The other limit is on the upper values of the vertical axis, which correspond to the direction of the adposition (prepositions). In other words, according to Greenberg, no languages can have prepositions and show some occurrence of relative sentences modifying nouns before head; they must be postpositional languages (in this part of the Universal). As can be seen, many languages are present in the rectangle that should be empty. Therefore, it does not seem that this part of the Universal is interesting from a typological point of view. Instead, we believe that an alternative and freer representation, such as the one shown in figure 21.58, is much more relevant.

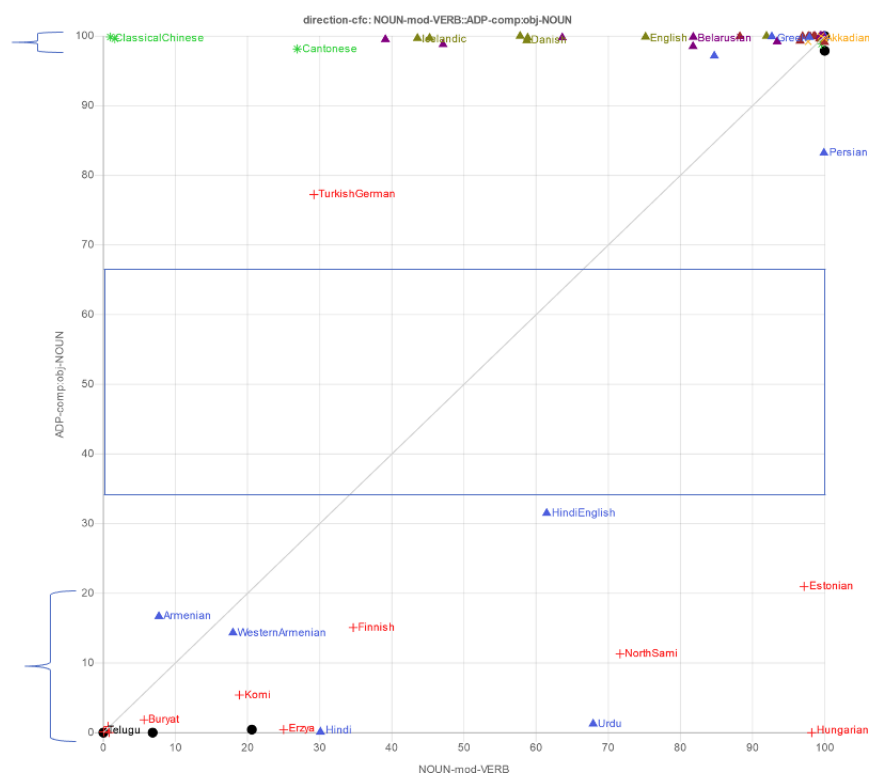


Figure 21.58: Verbs Modifying Nouns and Adposition Direction Proposal

In this case, we want to highlight the gap in the middle of the graph. This gap is due, in the first place, to the tendency of languages to be consistently prepositional or adpositional without exhibiting significant mixtures. On the other hand, it also seems that there tends to be a gap in the middle of the other axis, especially in the postpositional part, without allowing mixed occurrences. It is also important to note that, although it is clear that there are two solid poles (0-0 and 100-100) that correspond to consistencies of the two linguistic types, occurrences are found in the opposite corners (0-100 and 100-0).

Something different occurs with the second part of the Universal, represented in the figure 21.59.

In this case, we can see how Basque is the only language that goes against this formulation. We can also see some languages on the right-hand side that fall within the square. From an approach based on Typometrics, according to authors such as Sylvain Kahane or Kim Gerdes, these cases are not significant, given their proximity to the limit of the confirmation of the rule, their pattern that seems homogeneous and clear, and, in addition, we must remember that we are working with specific texts, which could slightly modify such a figure with another source. From a general point of view, it is evident that the behavior of the languages is totally different from the previous part

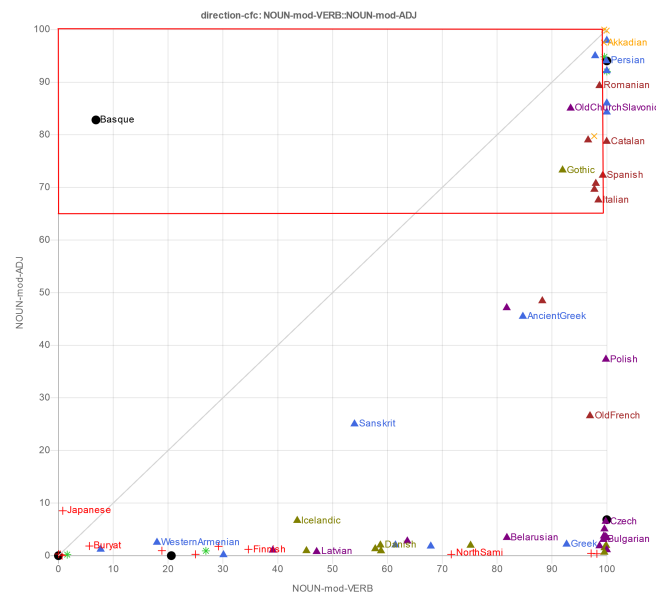


Figure 21.59: Verbs Modifying Nouns and Adjectives Modifying Nouns Direction Checking

of the Universal, and very clear restrictions are seen in the languages that place the adjectives behind the noun, only allowing the constructions that are treated in this Universal in an extreme postpositional way. It is striking to see how once we review languages with greater flexibility in the direction of the nominal modifier adjective or that direct adjective precedes modified nouns, there are occurrences on the entire axis, from 0 to 100.

The reading we propose of this Universal from a graphical point of view is that of an inverted L, as we point out in figure 21.60.

This proposed distribution highlights prominent and attractive restrictions (except for Basque and Sanskrit). This “L” shows us how languages that place verbs (in relative structures or not) noun modifiers inflexibly after the head show different values of the direction of the adjective modifying the noun. In contrast, languages that show occurrences of verb structures as noun modifiers in multiple directions do not offer such flexibility in the placement of adjectives, which are always prenominal. We are, therefore, before what appears to be a bidirectional implicature, in which the presence of one feature is associated with the other in whichever direction. Moreover, the restriction in the upper left corner is also very well observed: languages with a clear dominance of preceding modifier verbs have no occurrences of postnominal adjectives, and languages with postnominal adjectives have no occurrences of preceding modifier verbs.

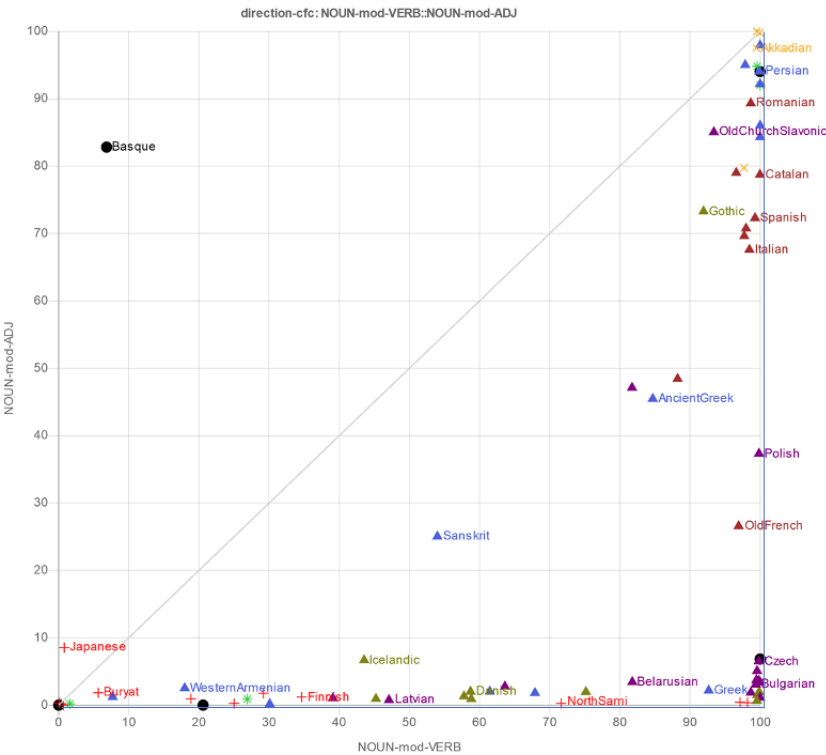


Figure 21.60: Verbs Modifying Nouns and Adjectives Modifying Nouns Direction Proposal

21.24.4 Trichotomous Expression Based on Grew-Match Data

Once reviewed in this visual and general way (verbs such as dependent are included as part of relative clause structures, but also other cases), we can review the quantitative data of the precise structure exposed by Greenberg thanks to Grew-Match, taking considering exclusively the cases of the relative clause. In this case, we should point out that the compliance of the Universal is 94%.

This high result is due to two reasons. First is the type of formulation of the Universal requires one of two options to be valid (or both simultaneously) to be considered valid. Secondly, one of the options that can contribute to the validity of the Universal (the one related to the direction of the adjective) has a predictive power of 89%.

In contrast, the first part of the Universal shows compliance values of 50%, which is of very little value for formulating typological claims. In addition, we may note that the languages exhibiting compliance with both parts (the last possibility described by Greenberg) is 44%. This figure is weighed down by the 50% of compliance described above.

Thus, these more precise data allow us to confirm that the more general approach previously shown in the Typometrics plots pointed to correct patterns and, therefore, this claim can be

extended to cases like the one described by Greenberg but also less restricted, something that gains in explanatory power and Universality. We offer the figure 21.61 as the realization of the GLU Model.

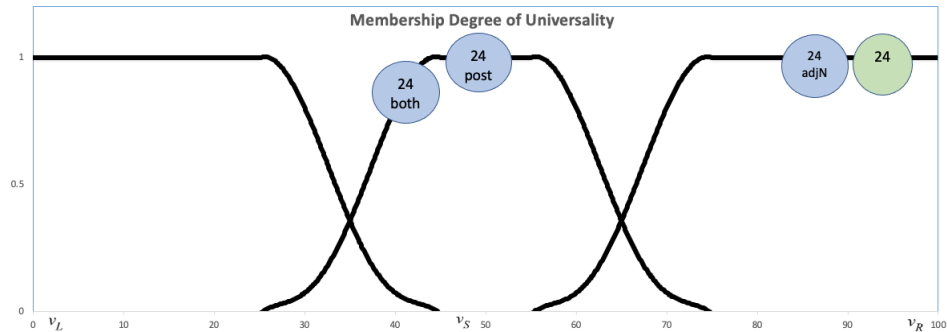


Figure 21.61: Trichotomous Expression for Universal 24

In this arrangement we offer, it is clear that the Universal formulated in general terms is a prototypical High-Universal, as well as the part referring to the direction of the nominal modifying adjective. On the other hand, the part referring to the direction of the adposition (and the Universal that takes into account the fulfillment of both parts at the same time) is a Mid-Universal with relatively central values, which implies that it is not attractive to contemplate such information as a predictor of linguistic types. Therefore, although one can keep the Universal as it is formulated because of its high predictive power, there is also the possibility of carrying out a reduction such as:

- If the relative expression precedes the noun as the only construction or as an alternate construction, then the adjective precedes the noun.

We can summarise our findings as in table 21.29.

Summary of characteristics of the Universal 24 analysis: 1. Typometrics: There are several languages with varying degrees of precedence to the noun as an alternative order in prepositional languages. 2. Typometrics: Prepositional languages show greater rigidity and behavior similar to postpositional languages. 3. Typometrics: Prepositional languages generally allow very few occurrences of variation in the order of nominal modification by verb structures. Postpositional languages show more heterogeneous behavior. 4. Typometrics: The direction of the adjective is a more reliable criterion for understanding the functioning of the languages of this Universal. Only Basque behaves anomalously. 5. Typometrics: Languages that place the modifying adjective before the noun show great flexibility in the direction of noun-modifying verbs. There is an inverted L-shaped distribution. 6. The original Universal is fulfilled in 94% of the cases. 7. The adjective direction part has a predictive power of 89%, and the adposition part has a predictive power of 50 8. The fulfillment of both parts simultaneously occurs in 44% of the cases. 9. Greenberg’s original Universal is a High-Universal, as is the part of the Universal relating to the address of the nominal modifier adjective. 10. The fulfillment of the postpositional part of the Universal or the fulfillment of both parts are Mid-Universals.

Table 21.29: Summary of Characteristics of the Universal 24 Analysis

21.25 Universal 25

UNIVERSAL 25: *If the pronominal object follows the verb, so does the nominal object.*

21.25.1 How to Check It in Typometrics

Technical Decisions: 1. Firstly, the Universal is analyzed with Typometrics (Gerdes et al., 2021). 2. We select the SUD annotation scheme because we can get results more comparable to the classical typological claims. 3. We choose the 2D Graph option. 4. We set a filter of utterances to avoid misleading data. 5. We select <i>direction – cfc</i> and <i>VERB – comp : obj – NOUN</i> for the x-Axis and <i>VERB – comp : obj – PRON</i> for the y-Axis, which is corresponding with the original Universal, the correlation between pronouns and nouns after the verb. 6. We also want to check how proper nouns behave, so we select <i>direction – cfc</i> and <i>VERB – comp : obj – PROP</i> for the x-Axis and maintain <i>VERB – comp : obj – PRON</i> for the y-Axis in the second graph. This graph would be equal to adding the word “proper” in the original Universal before the word “noun”. 7. Finally, we present an extra graph with one correlation that would be worth checking, related to the order of the nominal object placement. We select <i>direction – cfc</i> and <i>VERB – comp : obj – NOUN</i> for the x-Axis and <i>VERB – mod – ADV</i> for the y-Axis. This would be equal to claiming anything related to the position of the noun as a direct object and the position of an adverb modifying a noun, something that has been treated by comparing grammars in the latter bibliography as Dryer (2013). 8. Other possible correlations we would like to investigate are that of adjectives in predicative dependency and the order of adpositions. Therefore, we propose to create a graph with an X-axis of <i>VERB – comp : obj – NOUN</i> and y-Axis of <i>VERB – comp : pred – ADJ</i>; another with <i>VERB – comp : obj – PRON</i> and <i>VERB – comp : pred – ADJ</i>; another with <i>VERB – comp : obj – NOUN</i> and <i>ADP – comp : obj – NOUN</i>; and, one with <i>VERB – comp : obj – PRON</i> and <i>ADP – comp : obj – NOUN</i>.

Table 21.30: Technical Decisions for Checking Universal 25 in Typometrics

21.25.2 Formalization Proposed for Grew-Match

If we want to apply everything proposed in Typometrics in Grew-Match, we must formalize these syntactic relations as (21.361) - (21.372).

$$\begin{array}{l} \%25 - PN - before \\ pattern\{V[upos = VERB]; V - [comp : obj] - > PN; PN[upos = \\ PRON]; V >> PN\} \end{array} \quad (21.361)$$

(21.361) captures pronominal direct objects before the verb.

$$\begin{array}{l} \%25 - PN - after \\ pattern\{V[upos = VERB]; V - [comp : obj] - > PN; PN[upos = \\ PRON]; V << PN\} \end{array} \quad (21.362)$$

(21.362) captures pronominal direct objects after the verb.

$$\begin{array}{l} \%25 - N - before \\ pattern\{V[upos = VERB]; V - [comp : obj] - > N; N[upos = \\ NOUN|PROPN]; V >> N\} \end{array} \quad (21.363)$$

(21.363) captures nominal direct objects before the verb.

$$\begin{array}{l} \%25 - N - after \\ pattern\{V[upos = VERB]; V - [comp : obj] - > N; N[upos = \\ NOUN|PROPN]; V << N\} \end{array} \quad (21.364)$$

(21.364) captures nominal direct objects after the verb.

$$\begin{array}{l} \%25 - ADJ - predicative - before \\ pattern\{V[upos = VERB|AUX]; V - [comp : pred] - > \\ ADJ; ADJ[upos = ADJ]; V >> ADJ\} \end{array} \quad (21.365)$$

(21.365) captures adjectival predicatives before the verb.

$$\begin{array}{l} \%25 - ADJ - predicative - after \\ pattern\{V[upos = VERB|AUX]; V - [comp : pred] - > \\ ADJ; ADJ[upos = ADJ]; V << ADJ\} \end{array} \quad (21.366)$$

(21.366) captures adjectival predicatives after the verb.

$$\begin{array}{l} \%25 - ADP - before \\ pattern\{V[upos = VERB]; V - [comp : obj] - > ADP; ADP[upos = \\ ADP]; V >> ADP\} \end{array} \quad (21.367)$$

(21.367) captures depending prepositions before the head verb.

$$\begin{array}{l} \%25 - ADP - after \\ pattern\{V[upos = VERB]; V - [comp : obj] - > ADP; ADP[upos = \\ ADP]; V << ADP\} \end{array} \quad (21.368)$$

(21.368) captures depending prepositions after the head verb.

$$\begin{array}{l} \%25 - ADV - before \\ pattern\{V[upos = VERB]; V - [mod] - > ADV; ADV[upos = \\ ADV]; V >> ADV\} \end{array} \quad (21.369)$$

(21.369) captures adverbial modifiers before the verb.

$$\begin{array}{l} \%25 - ADV - after \\ pattern\{V[upos = VERB]; V - [mod] - > ADV; ADV[upos = \\ ADV]; V << ADV\} \end{array} \quad (21.370)$$

(21.370) captures adverbial modifiers after the verb.

$$\begin{array}{l} \%25 - PN - before - ACC \\ pattern\{V[upos = VERB]; V - [comp : obj] - > PN; PN[upos = \\ PRON, Case = Acc]; V >> PN\} \end{array} \quad (21.371)$$

(21.371) captures pronominal direct objects with the accusative case before the verb.

$$\begin{array}{l} \%25 - PN - after - ACC \\ pattern\{V[upos = VERB]; V - [comp : obj] - > PN; PN[upos = \\ PRON, Case = Acc]; V << PN\} \end{array} \quad (21.372)$$

(21.372) captures pronominal direct objects with accusative case after the verb.

The first two patterns (21.361) and (21.362) correspond to the location of pronouns functioning as direct objects, before or after the verb on which they depend, thanks to the dependency $V - [comp : obj] - > PN$. The following two patterns (21.364) and (21.363) refer to the location of common and proper nouns functioning as direct objects (here we decided to treat them together because it is the only metric we cannot obtain in Typometrics), before or after the verb they depend on, thanks to the dependency $V - [comp : obj] - > N$. The following two patterns (21.365) and (21.366) correspond to the location of the predicative, something we have already commented that deserves a separate, more detailed future study, thanks to the dependency $V - [comp : pred] - > ADJ$. Subsequently, we find two patterns (21.367) and (21.368) related to the position of the adposition about the noun, one of the fundamental relations for Greenberg $ADP - [comp : obj] - > N$. The following two patterns (21.369) and (21.370) correspond to the last variable checked in Typometrics related to one of Dryer's object patterners, the position of the adverb as verbal modifier $V - [mod] - > ADV$. The last two patterns (21.371) and (21.372) formalize a new metric that we only obtain by Grew-Match since it is very precise. This corresponds to adding the distinction of valuing the position of pronouns that function as direct objects and have accusative cases. This allows us to understand whether case marking can influence this Universal as is suspected to always occur with the free word order and languages with more marked cases. To do this, what is done is to go back to the patterns (21.361) and (21.362) about pronoun and verb order and change the *PRON* value to $[upos = PRON, Case = Acc]$, a fact that filters out only the cases in which that pronoun is in the accusative case.

21.25.3 Graphs from Typometrics

Verifying this Universal by Typometrics is easy and straightforward, as shown in figure 21.62.

We should expect not to find occurrences in the upper part of the vertical axis (pronominal object after the verb) that are on the left (nominal object before the verb) since, according to Greenberg, the nominal object must also be after the verb (on the right). As can be seen, indeed, we can see how the three possible corners of the graph contain different languages and a much stronger restriction appears in the area designated by Greenberg as non-Universal. We only find two languages that are Semitic/Afroasiatic, a fact that indicates that this may be a unique characteristic of this group.

In addition, we believe it is also illustrative to show the graph highlighting the lower right triangular area as in figure 21.63.

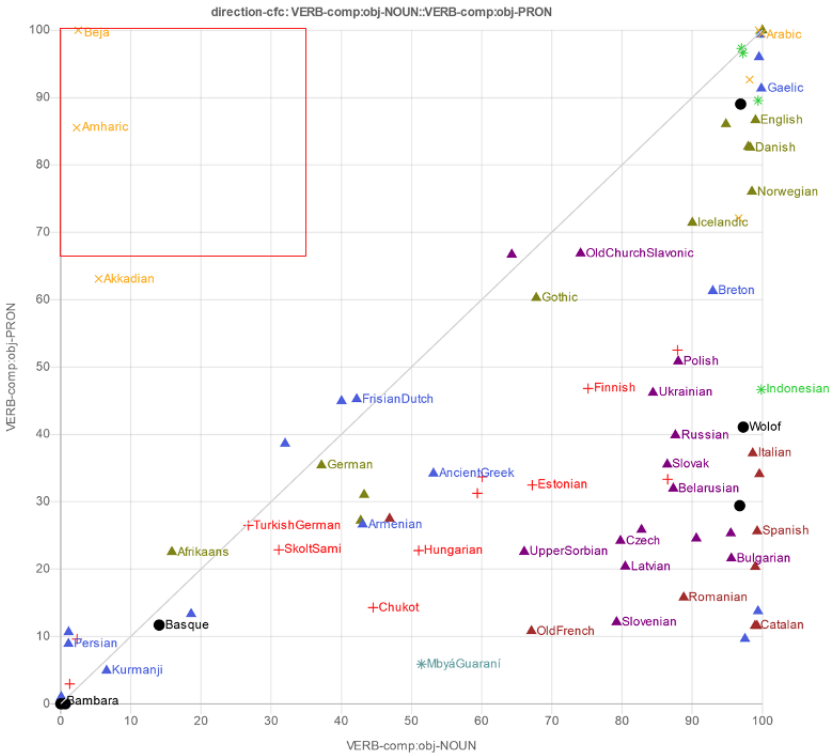


Figure 21.62: Nominal (Common) and Pronominal Direct Objects Direction Checking

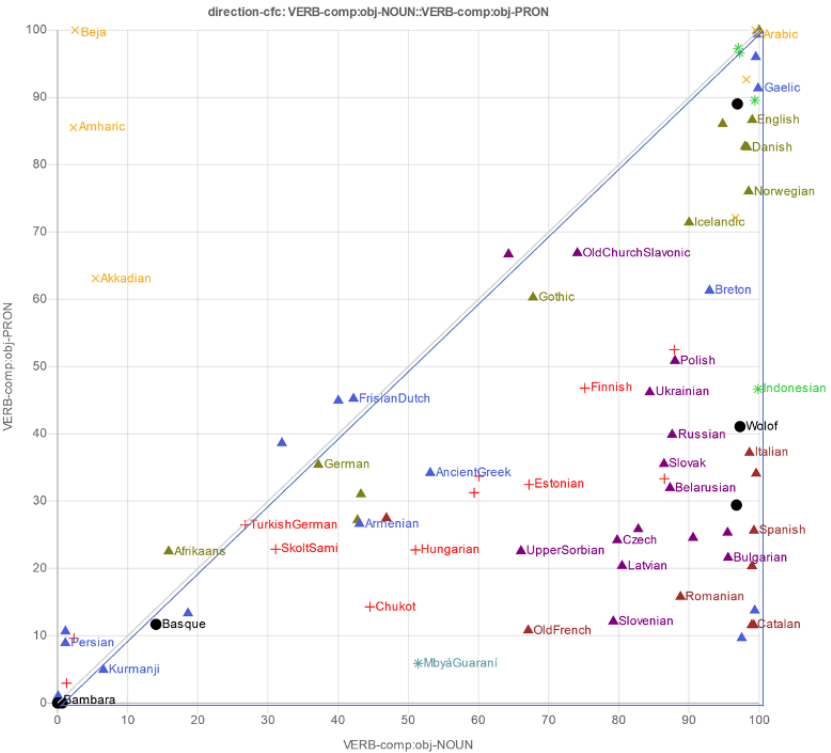


Figure 21.63: Nominal (Common) and Pronominal Direct Objects Direction Proposal

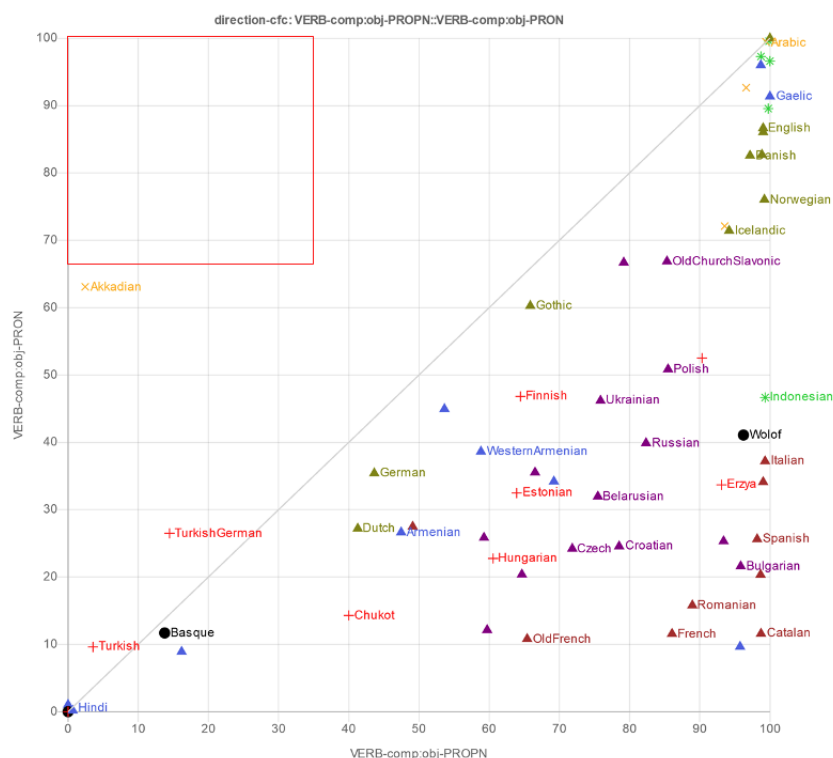


Figure 21.64: Nominal (Proper) and Pronominal Direct Objects Direction Checking

This triangular arrangement quantitatively shows us that the vast majority of the world's languages always exhibit a higher proportion of pronominal direct objects preceding the verb than of nominal direct objects, no matter what this proportion is. This may give us an idea of the tendency of direct object pronouns to be preverbal, whereas nominal direct objects do not have this clear tendency.

In addition, we can be more precise and differentiate and analyze the behavior of proper nouns (compared to the common ones shown above). Such a change in the precision of the Universal can be seen in figure 21.64.

In this case, we see that no language appears within the square that we propose to delimit the zone designated by Greenberg as non-Universal. However, it should be noted that the two languages that appeared have not moved to any place but have disappeared. This assumes that the absence of the languages previously present in the exception zone is because there are not enough occurrences of the designated type in the corpus analyzed to show conclusive results.

Once the exception zone is, it is equally interesting to look at the overall distribution of languages in the graph and compare it with the distribution of common names, as shown in figure 21.65.

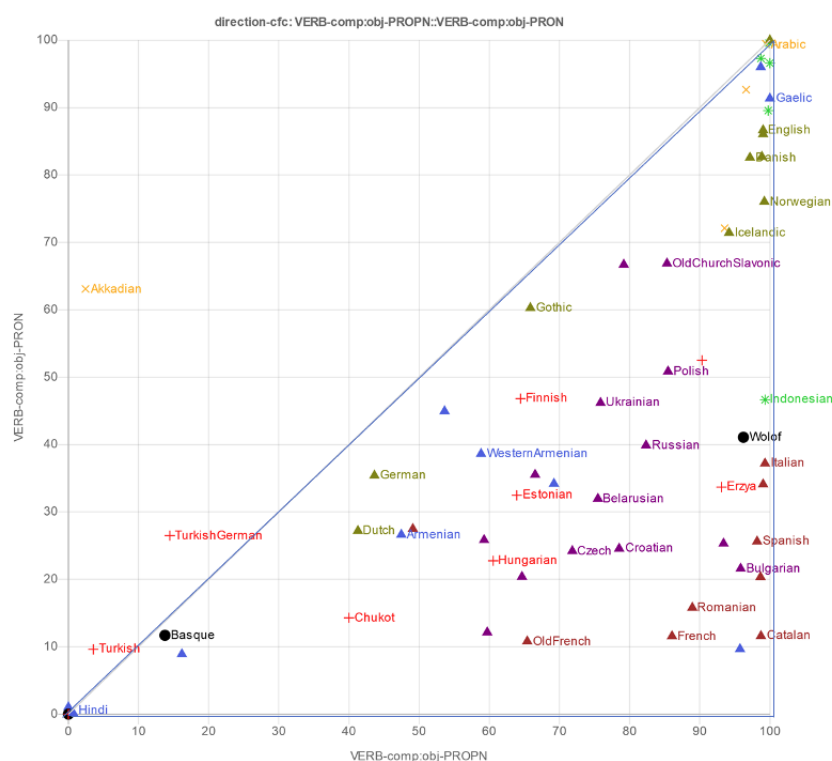


Figure 21.65: Nominal (Proper) and Pronominal Direct Objects Direction Proposal

In general, we must comment that the placement of languages in the lower right triangle is still respected. Therefore, we draw the conclusion that the formulated Universal serves for both common nouns and proper nouns. In addition, we see a tendency for the languages to be located even further to the right, implying that fewer languages are on the border, separating one triangle from the other. In other words, this triangle is much more representative and clearer if possible.

21.25.4 Trichotomous Expression Based on Grew-Match Data

If we want to check this data by Grew-Match and categories, we obtain:

1. A compliance with the Universal of exactly 94%.
2. We also wanted to check the relationship of nouns and pronouns with the direction of prepositional complements. In this case, the preceding pronominal object predicts 85% of the cases of prepositional complements.
3. The preceding nominal object predicts 86% of the cases of preceding prepositional complements.

This information, transferred to the trichotomous expression, is represented in figure 21.66:

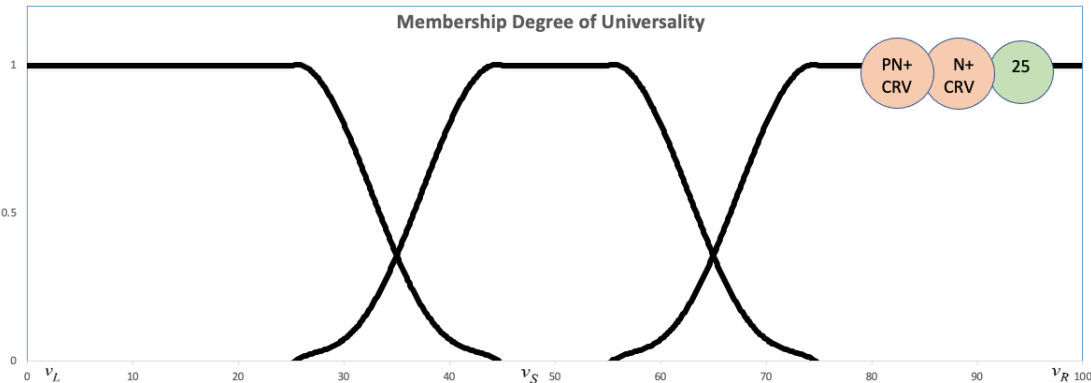


Figure 21.66: Trichotomous Expression for Universal 25

In the three cases tested (Greenberg’s original Universal and the two tests with verbal regimes), we are dealing with prototypical high Universals. Therefore, both the Universal formulated by Greenberg and our extension proposals are fully valid and of great interest when predicting the behavior of languages in general. We can summarise our findings as in table 21.31.

Summary of characteristics of the Universal 25 analysis:
1. Typometrics: Some Semitic languages are an exception to the Universal, but Greenberg’s trend is observed.
2. Typometrics: There is a tendency for languages to show a higher proportion of nominal occurrences than pronominal direct objects.
3. Typometrics: The behavior with proper nouns is very similar to the above, although with no exceptions.
4. The original Universal has a compliance of 94%.
5. The relation of pronouns and nouns to the direction of prepositional complements is 85% correct. In the case of exclusively nominal objects, it is 86%.
6. Greenberg’s original Universal and the proposed linking with verbal regime complements are high-Universals.

Table 21.31: Summary of Characteristics of the Universal 25 Analysis

21.26 Universal 26

UNIVERSAL 26: *If a language has discontinuous affixes, it always has either prefixing, suffixing, or both.*

21.26.1 Current Limitation: Data no Available

In this Universal case, again, we have the problem of not having specific information on the presence and behavior of affixes in the languages present in Universal Dependencies. In the case of discontinuous affixes, an affix that manifests itself in different parts of the same word, it is much more important to have precise and detailed information, which we do not have. We believe that it is especially relevant and helpful to use interglossae to collect information related to affixes and similar elements in an organized, coherent, and visual way.

21.27 Universal 27

UNIVERSAL 27: *If a language is exclusively suffixing, it is postpositional; if exclusively prefixing, it is prepositional.*

21.27.1 Current Limitation: Data no Available

When working with suffixing and prefixing, this Universal is again dealing with affixes. Therefore, we also cannot know these elements' behavior in the world's languages through UD.

21.28 Universal 28

UNIVERSAL 28: *If both the derivation and inflection follow the root or precede the root, the derivation is always between the root and the inflection.*

21.28.1 Current Limitation: Data no Available

Again, in Universal 28, we work with elements such as derivation and inflection. There is no information related to this. Again, we believe this Universal could be easily checked if we used an annotation through interglossae, which allows us to know all the elements inside a word and establish a clear correlation between that language and ours.

21.29 Universal 29

UNIVERSAL 29: *If a language has inflection, it always has derivation.*

21.29.1 Current Limitation: Data no Available

This Universal, like Universal 28 (21.28), is based on the opposition between inflection and derivation, which we cannot certify with the data available in the analyzed treebanks.

21.30 Universal 30

UNIVERSAL 30: *If the verb has categories of person-number or if it has categories of gender, it always has tense-mode categories.*

21.30.1 Extracting the Information in the Treebank Hub Page

This Universal does not ask for any particular dependency, just the presence of a feature or not. This can be done superficially with the Treebank hub page in Universal Dependencies. However,

we must note that the “categories of person-number” part of the Universal is ambiguous. We cannot be sure if he states that both must occur or just one of them. Alternatively, maybe it implies that if one exists, the other too (biconditional). We find it more reasonable to consider that the use of “-” in this case is equal to “&”.

We must go to each corpus’s treebank hub page and select “POS Tags> VERB”. Then, we must check the morphology part. Check the features cited with Verbs in the corpus selected (person, number, gender, tense, mode). If we want to confirm the Universal, all languages that show person and number, or gender will also exhibit tense and mode.

For example, in Spanish (AnCora), we see that the Verb part of speech occurs with the features VerbForm, Number, Tense, Mood, Person, and Gender. So here we find an example where we have person, number, and gender. We also find tension and mood. So, all the cited features are present in this language (or corpus).

21.30.2 Formalization Proposed for Grew-Match

Subsequently, we can propose a more systematized check (although it requires extra knowledge beyond the check on treebank information) using Grew-Match. The check to be made is simple: we must search for occurrences in a treebank that respond to the feature (categories in Greenberg’s terms) announced in the Universal. To check the Universal, we must always formalize verbs and check the features of person, number, gender, tense, and mode. We propose (21.373) - (21.377).

$\%30 - V - Person$ $pattern\{V[upos = VERB, Person]\}$	(21.373)

(21.373) captures verbs with any person feature activated.

$\%30 - V - Number$ $pattern\{V[upos = VERB, Number]\}$	(21.374)

(21.374) captures verbs with any number feature activated.

$\%30 - V - Gender$ $pattern\{V[upos = VERB, Gender]\}$	(21.375)

(21.375) captures verbs with any gender feature activated.

$$\begin{array}{l} \%30 - V - Tense \\ pattern\{V[upos = VERB, Tense]\} \end{array} \quad (21.376)$$

(21.376) captures verbs with any tense feature activated.

$$\begin{array}{l} \%30 - V - Mood \\ pattern\{V[upos = VERB, Mood]\} \end{array} \quad (21.377)$$

(21.377) captures verbs with any mood feature activated.

However, to make a real check of Greenberg's Universal and not limit ourselves to the simple ratification of his words, we propose going further to see if we can extend the results. The extension can be done by checking more features that could affect verbs. In this case, we propose to check all possible verbal features not mentioned with the Universal, as voice, case, animacy, and polarity are in patterns (21.378) - (21.381).

$$\begin{array}{l} \%30 - V - Voice \\ pattern\{V[upos = VERB, Voice]\} \end{array} \quad (21.378)$$

(21.378) captures verbs with any voice feature activated.

$$\begin{array}{l} \%30 - V - Case \\ pattern\{V[upos = VERB, Case]\} \end{array} \quad (21.379)$$

(21.379) captures verbs with any case feature activated.

$$\begin{array}{l} \%30 - V - Animacy \\ pattern\{V[upos = VERB, Animacy]\} \end{array} \quad (21.380)$$

(21.380) captures verbs with any animacy feature activated.

$$\begin{array}{l} \%30 - V - Polarity \\ pattern\{V[upos = VERB, Polarity]\} \end{array} \quad (21.381)$$

(21.381) captures verbs with any polarity feature activated.

By extending our study to more features, some of them much more transversal, such as “case”, we try to find an invariant behavior in the languages of the world, whatever the feature, that allows us to establish a classification, perhaps, as an implicational hierarchy. However, we must specify

that the Universal specifies that the same verb is one that must contain all these categories at the same time. Therefore, it is necessary to make a much more precise check (that will offer fewer occurrences), showing all the possible co-occurrences of features in the verbs of the languages of the world (and possible incompatibilities or exclusions). To do so, we must combine all the features announced by Greenberg among them. In order to be more representative, we do it progressively, starting with combinations of pairs (21.382) - (21.391); then moving on to triplets (21.392) - (21.401); then quadruples (21.402) - (21.406); and then all five features at the same time, as show in the pattern (21.407).

$$\begin{array}{l} \%30 - V - Person - number \\ pattern\{V[upos = VERB, Person, Number]\} \end{array} \quad (21.382)$$

(21.382) captures verbs with any person and number features activated.

$$\begin{array}{l} \%30 - V - Person - gender \\ pattern\{V[upos = VERB, Person, Gender]\} \end{array} \quad (21.383)$$

(21.383) captures verbs with any person and gender features activated.

$$\begin{array}{l} \%30 - V - Person - tense \\ pattern\{V[upos = VERB, Person, Tense]\} \end{array} \quad (21.384)$$

(21.384) captures verbs with any person and tense features activated.

$$\begin{array}{l} \%30 - V - Person - mood \\ pattern\{V[upos = VERB, Person, Mood]\} \end{array} \quad (21.385)$$

(21.385) captures verbs with any person and mood features activated.

$$\begin{array}{l} \%30 - V - Number - Gender \\ pattern\{V[upos = VERB, Number, Gender]\} \end{array} \quad (21.386)$$

(21.386) captures verbs with any number and gender features activated.

$$\begin{array}{l} \%30 - V - Number - Tense \\ pattern\{V[upos = VERB, Number, Tense]\} \end{array} \quad (21.387)$$

(21.387) captures verbs with any number and tense features activated.

$$\begin{array}{l} \%30 - V - Number - Mood \\ pattern\{V[upos = VERB, Number, Mood]\} \end{array} \quad (21.388)$$

(21.388) captures verbs with any number and mood features activated.

$$\begin{array}{l} \%30 - V - Gender - Tense \\ pattern\{V[upos = VERB, Gender, Tense]\} \end{array} \quad (21.389)$$

(21.389) captures verbs with any gender and tense features activated.

$$\begin{array}{l} \%30 - V - Gender - Mood \\ pattern\{V[upos = VERB, Gender, Mood]\} \end{array} \quad (21.390)$$

(21.390) captures verbs with any gender and mood features activated.

$$\begin{array}{l} \%30 - V - Tense - Mood \\ pattern\{V[upos = VERB, Tense, Mood]\} \end{array} \quad (21.391)$$

(21.391) captures verbs with any tense and mood features activated.

$$\begin{array}{l} \%30 - V - Person - Number - Gender \\ pattern\{V[upos = VERB, Person, Number, Gender]\} \end{array} \quad (21.392)$$

(21.392) captures verbs with any person, number, and gender features activated.

$$\begin{array}{l} \%30 - V - Person - Number - Tense \\ pattern\{V[upos = VERB, Person, Number, Tense]\} \end{array} \quad (21.393)$$

(21.393) captures verbs with any person, number, and tense features activated.

$$\begin{array}{l} \%30 - V - Person - Number - Mood \\ pattern\{V[upos = VERB, Person, Number, Mood]\} \end{array} \quad (21.394)$$

(21.394) captures verbs with any person, number, and mood features activated.

$$\begin{array}{l} \%30 - V - Person - Gender - Tense \\ pattern\{V[upos = VERB, Person, Gender, Tense]\} \end{array} \quad (21.395)$$

(21.395) captures verbs with any person, gender, and tense features activated.

$$\begin{array}{l} \%30 - V - Person - Gender - Mood \\ pattern\{V[upos = VERB, Person, Gender, Mood]\} \end{array} \quad (21.396)$$

(21.396) captures verbs with any person, gender, and mood features activated.

$$\begin{array}{l} \%30 - V - Person - Tense - Mood \\ pattern\{V[upos = VERB, Person, Tense, Mood]\} \end{array} \quad (21.397)$$

(21.397) captures verbs with any person, tense, and mood features activated.

$$\begin{array}{l} \%30 - V - Number - Gender - Tense \\ pattern\{V[upos = VERB, Number, Gender, Tense]\} \end{array} \quad (21.398)$$

(21.398) captures verbs with any number, gender, and tense features activated.

$$\begin{array}{l} \%30 - V - Number - Gender - Mood \\ pattern\{V[upos = VERB, Number, Gender, Mood]\} \end{array} \quad (21.399)$$

(21.399) captures verbs with any number, gender, and mood features activated.

$$\begin{array}{l} \%30 - V - Number - Tense - Mood \\ pattern\{V[upos = VERB, Number, Tense, Mood]\} \end{array} \quad (21.400)$$

(21.400) captures verbs with any number, tense, and mood features activated.

$$\begin{array}{l} \%30 - V - Gender - Tense - Mood \\ pattern\{V[upos = VERB, Gender, Tense, Mood]\} \end{array} \quad (21.401)$$

(21.401) captures verbs with any gender, tense, and mood features activated.

$$\begin{array}{l} \%30 - V - Person - Number - Gender - Tense \\ pattern\{V[upos = VERB, Person, Number, Gender, Tense]\} \end{array} \quad (21.402)$$

(21.402) captures verbs with any person, number, gender, and tense features activated.

$$\begin{array}{l} \%30 - V - Person - Number - Gender - Mood \\ pattern\{V[upos = VERB, Person, Number, Gender, Mood]\} \end{array} \quad (21.403)$$

(21.403) captures verbs with any person, number, gender, and mood features activated.

$$\begin{array}{l} \%30 - V - Person - Number - Tense - Mood \\ pattern\{V[upos = VERB, Person, Number, Tense, Mood]\} \end{array} \quad (21.404)$$

(21.404) captures verbs with any person, number, tense, and mood features activated.

$$\begin{array}{l} \%30 - V - Person - Gender - Tense - Mood \\ pattern\{V[upos = VERB, Person, Gender, Tense, Mood]\} \end{array} \quad (21.405)$$

(21.405) captures verbs with any person, gender, tense, and mood features activated.

$$\begin{array}{l} \%30 - V - Number - Gender - Tense - Mood \\ pattern\{V[upos = VERB, Number, Gender, Tense, Mood]\} \end{array} \quad (21.406)$$

(21.406) captures verbs with any number, gender, tense, and mood features activated.

$$\begin{array}{l} \%30 - V - Person - Number - Gender - Tense - Mood \\ pattern\{V[upos = VERB, Person, Number, Gender, \\ Tense, Mood]\} \end{array} \quad (21.407)$$

(21.407) captures verbs with any person, number, gender, tense, and mood features activated.

21.30.3 Trichotomous Expression Based on Grew-Match Data

This formulation of Greenberg's is again motivated by a description of the behavior of the languages of the world in general by its formulation, and we note that it is valid for 44% of the languages available to us.

The trichotomous expression that we can generate from this Universal is shown in figure 21.67.

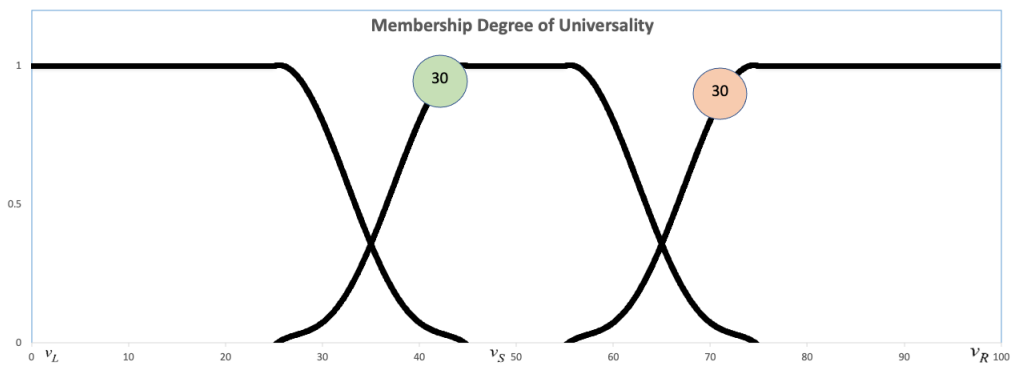


Figure 21.67: Trichotomous Expression for Universal 30

The results show a Mid-Universal. This Mid-Universal is due to a problem with the conception of the mood-tense formulation in Greenberg, related to the annotation in UD. In many languages, some moods show no distinction between tenses since there is only one. This is the case of imperative in Spanish, for example. Thus, technically, in Spanish, it is possible to find verbs with person-number categories but not tense (only mode). This fact causes us to find occurrences that Greenberg marked as not Universal. If we are strict, we must consider that the Universal is not fulfilled. For this reason, we recommend just formulating it in terms of mood (or mode) and not considering both mood and tense. We can summarise our findings as in table 21.32.

Summary of characteristics of the Universal 30 analysis: 1. The compliance with Greenberg’s original Universal is 44%. 2. The low result is due to an absence of tense-mood in quite a few languages in some specific moods without annotated tense. 3. Greenberg’s Universal is a Mid-Universal. 4. A proposed modification of the Universal to consider only mood makes it a High-Universal.

Table 21.32: Summary of Characteristics of the Universal 30 Analysis

21.31 Universal 31

UNIVERSAL 31: *If either the subject or object noun agrees with the verb in gender, then the adjective always agrees with the noun in gender.*

21.31.1 Formalization Proposed for Grew-Match

In order to check this Universal, we must check, on the one hand, whether there is gender agreement between subjects or objects (it does not matter which) and the verb. On the other hand, we must check if there is gender agreement between a noun and an adjective that modifies it. The implication formulated assumes that there are no languages in which there is gender agreement in the subject or object, and the same agreement does not exist between an adjective and a noun. In order words, if the Universal is correct, there are only three possibilities valid out of four:

1. There is no agreement at all in both cases.

2. There is agreement only in noun-adjective relation.
3. There is an agreement in both.

The first additional distinction that we are going to make and that Greenberg does not establish is precisely between subjects and objects. We want to check whether both elements contribute in the same way to the Universal. Therefore, we propose three possible formalizations: one that takes into account both at the same time, one for subjects, and one for objects.

In the proposed formalizations, we have always included a double negative filter (without) that allows us to discard two structures from our findings. We want to avoid occurrences in which several nouns are joined by a conjunction (first without) or a comma (second without). This will allow us to avoid occurrences in which there is a pair of nouns labeled with different genders and then a verb showing a different value than the selected word, for example. However, we must warn that what cannot be solved is when the speaker decides to make the concord to the gender or number in words such as “most” (more problematic for number than for gender). We cannot compute this because we do not have a list of all the constructions that could cause this in each language (it is also challenging to make such a list in even one language). However, the numbers of such cases are marginal and equal in the languages, so we do not foresee any influence on the dominant trend in languages. We can restrict special cases on the syntactic level, but you have no such information on the semantic level.

Since some treebanks may have tagging problems where something is labeled as masculine when it is not gendered, we propose a combination of all logically possible options to understand when there is agreement. This mode allows us to see whether there is agreement between the direct object and the verb in a language. For example, there is no such agreement in this language if there are many occurrences of the pattern in which a verb is masculine, and the object is feminine. Obviously, there is no such agreement if the verb (or any other element) does not have the active feature of agreement.

First, we present the formalization that takes into account, at the same time, the agreement of subjects or objects with the verb. For this purpose, we always activate the verb and noun elements. We ask that the dependency relation be subject or object. We add negative filters to avoid having nouns with different genders that may influence the final agreement (for example, in “Juan y Laura son mexicanos”, in Spanish, we would find a case of bad agreement between Laura (fem) and mexicanos (masc) if we do not know what is happening). In each of the patterns, we

configure the features of the two elements differently, with all the possibilities: that the verb and the noun have no gender, that the verb is masculine and the noun has no gender, that the verb is feminine and the noun has no gender, that the verb is masculine and the noun is feminine, that the verb is feminine and the noun is feminine, etc. In the case of verifying the presence of the gender feature in one of the two words and the absence in the other, it will be sufficient to verify that there are occurrences with a given gender to certify that there is no such agreement. For example, in the first pattern, we add the the !Gender feature of the verb and the noun (subject or object) means that neither of the two words has the gender feature active. In the next one, for example, we look for cases where the verb still has no gender value, but some nouns (subjects or objects) have a feminine gender. In the following case, we change the variable and look for verbs with the feminine gender value and nouns with no gender. In the following patterns, we play with more possible agreements (two feminine words, two masculine words) and cases that would indicate a lack of agreement but have suffered bad tagging (one word is masculine and the other feminine). The formalizations closer to Greenberg's Universal are as follows in (21.408) - (21.414).

$$\begin{array}{l}
 \%31 - Subj - Obj - Nagree - both \\
 pattern\{V[upos = VERB,!Gender]; V - [comp : obj|subj]- > \\
 N; N[upos = NOUN|PROPN,!Gender]\} \\
 without\{N2[upos = ADJ|ADV|INTJ|NOUN|PROPN| \\
 VERB|ADP|AUX|CCONJ|DET|NUM|PART|PRON| \\
 SCONJ|SYM]; N - [conj]- > N2\} \\
 without\{N2[upos = NOUN|PROPN]; N - [appos]- > N2\}
 \end{array}
 \tag{21.408}$$

(21.408) captures nominal subjects or objects without gender and verbs without gender.

$$\begin{array}{l}
 \%31 - Subj - Obj - Nagree - V \\
 pattern\{V[upos = VERB,!Gender]; V - [comp : obj|subj]- > \\
 N; N[upos = NOUN|PROPN, Gender = Fem]\} \\
 without\{N2[upos = ADJ|ADV|INTJ|NOUN|PROPN| \\
 VERB|ADP|AUX|CCONJ|DET|NUM|PART|PRON| \\
 SCONJ|SYM]; N - [conj]- > N2\} \\
 without\{N2[upos = NOUN|PROPN]; N - [appos]- > N2\}
 \end{array}
 \tag{21.409}$$

(21.409) captures nominal subjects or objects with feminine and verbs without gender.

$$\begin{aligned}
&\%31 - \textit{Subj} - \textit{Obj} - \textit{Nagree} - N \\
&\textit{pattern}\{V[\textit{upos} = \textit{VERB}, \textit{Gender} = \textit{Fem}]; V - [\textit{comp} : \textit{obj} | \textit{subj}] - > \\
&N; N[\textit{upos} = \textit{NOUN} | \textit{PROPN}, !\textit{Gender}]\} \\
&\textit{without}\{N2[\textit{upos} = \textit{ADJ} | \textit{ADV} | \textit{INTJ} | \textit{NOUN} | \textit{PROPN} | \\
&\textit{VERB} | \textit{ADP} | \textit{AUX} | \textit{CCONJ} | \textit{DET} | \textit{NUM} | \textit{PART} | \textit{PRON} | \\
&\textit{SCONJ} | \textit{SYM}]; N - [\textit{conj}] - > N2\} \\
&\textit{without}\{N2[\textit{upos} = \textit{NOUN} | \textit{PROPN}]; N - [\textit{appos}] - > N2\}
\end{aligned} \tag{21.410}$$

(21.410) captures nominal subjects or objects without gender and verbs with feminine.

$$\begin{aligned}
&\%31 - \textit{Subj} - \textit{Obj} - \textit{MascFem} \\
&\textit{pattern}\{V[\textit{upos} = \textit{VERB}, \textit{Gender} = \textit{Masc}]; V - [\textit{comp} : \textit{obj} | \textit{subj}] - > \\
&N; N[\textit{upos} = \textit{NOUN} | \textit{PROPN}, \textit{Gender} = \textit{Fem}]\} \\
&\textit{without}\{N2[\textit{upos} = \textit{ADJ} | \textit{ADV} | \textit{INTJ} | \textit{NOUN} | \textit{PROPN} | \\
&\textit{VERB} | \textit{ADP} | \textit{AUX} | \textit{CCONJ} | \textit{DET} | \textit{NUM} | \textit{PART} | \textit{PRON} | \\
&\textit{SCONJ} | \textit{SYM}]; N - [\textit{conj}] - > N2\} \\
&\textit{without}\{N2[\textit{upos} = \textit{NOUN} | \textit{PROPN}]; N - [\textit{appos}] - > N2\}
\end{aligned} \tag{21.411}$$

(21.411) captures nominal subjects or objects with feminine gender and verbs with masculine gender.

$$\begin{aligned}
&\%31 - \textit{Subj} - \textit{Obj} - \textit{FemMasc} \\
&\textit{pattern}\{V[\textit{upos} = \textit{VERB}, \textit{Gender} = \textit{Fem}]; V - [\textit{comp} : \textit{obj} | \textit{subj}] - > \\
&N; N[\textit{upos} = \textit{NOUN} | \textit{PROPN}, \textit{Gender} = \textit{Masc}]\} \\
&\textit{without}\{N2[\textit{upos} = \textit{ADJ} | \textit{ADV} | \textit{INTJ} | \textit{NOUN} | \textit{PROPN} | \\
&\textit{VERB} | \textit{ADP} | \textit{AUX} | \textit{CCONJ} | \textit{DET} | \textit{NUM} | \textit{PART} | \textit{PRON} | \\
&\textit{SCONJ} | \textit{SYM}]; N - [\textit{conj}] - > N2\} \\
&\textit{without}\{N2[\textit{upos} = \textit{NOUN} | \textit{PROPN}]; N - [\textit{appos}] - > N2\}
\end{aligned} \tag{21.412}$$

(21.412) captures nominal subjects or objects with masculine gender and verbs with feminine gender.

$$\begin{aligned}
&\%31 - \textit{Subj} - \textit{Obj} - \textit{MascMasc} \\
&\textit{pattern}\{V[\textit{upos} = \textit{VERB}, \textit{Gender} = \textit{Masc}]; V - [\textit{comp} : \textit{obj} | \textit{subj}] - > \\
&N; N[\textit{upos} = \textit{NOUN} | \textit{PROPN}, \textit{Gender} = \textit{Masc}]\} \\
&\textit{without}\{N2[\textit{upos} = \textit{ADJ} | \textit{ADV} | \textit{INTJ} | \textit{NOUN} | \textit{PROPN} | \\
&\textit{VERB} | \textit{ADP} | \textit{AUX} | \textit{CCONJ} | \textit{DET} | \textit{NUM} | \textit{PART} | \textit{PRON} | \\
&\textit{SCONJ} | \textit{SYM}]; N - [\textit{conj}] - > N2\} \\
&\textit{without}\{N2[\textit{upos} = \textit{NOUN} | \textit{PROPN}]; N - [\textit{appos}] - > N2\}
\end{aligned} \tag{21.413}$$

(21.413) captures nominal subjects or objects with masculine gender and verbs with masculine gender.

$$\begin{aligned}
&\%31 - \textit{Subj} - \textit{Obj} - \textit{FemFem} \\
&\textit{pattern}\{V[\textit{upos} = \textit{VERB}, \textit{Gender} = \textit{Fem}]; V - [\textit{comp} : \textit{obj} | \textit{subj}] - > \\
&N; N[\textit{upos} = \textit{NOUN} | \textit{PROPN}, \textit{Gender} = \textit{Fem}]\} \\
&\textit{without}\{N2[\textit{upos} = \textit{ADJ} | \textit{ADV} | \textit{INTJ} | \textit{NOUN} | \textit{PROPN} | \\
&\textit{VERB} | \textit{ADP} | \textit{AUX} | \textit{CCONJ} | \textit{DET} | \textit{NUM} | \textit{PART} | \textit{PRON} | \\
&\textit{SCONJ} | \textit{SYM}]; N - [\textit{conj}] - > N2\} \\
&\textit{without}\{N2[\textit{upos} = \textit{NOUN} | \textit{PROPN}]; N - [\textit{appos}] - > N2\}
\end{aligned} \tag{21.414}$$

(21.414) captures nominal subjects or objects with feminine gender and verbs with feminine gender.

In short, in this case, we can see one of the limitations of the corpus approach that does not appear in armchair methodology, in which commutations can be carried out to see if there is a change of gender or number of more elements. In this case, the only criterion to be followed is the one proposed. If we find many feminine subjects with feminine verbs and do not find occurrences of feminine subjects with masculine or without gender, for example, there is an agreement between these two elements.

Subsequently, we must check whether the language has occurrences of agreement between nouns and adjectives that modify them. In these cases, we repeat the combinations expressed above between nouns and verbs and apply them to the modification dependency between nouns and adjectives. This can be done using the patterns (21.415) - (21.421).

$$\begin{array}{l}
\%31 - AdjN - Nagree - both \\
pattern\{N[upos = NOUN|PROPN,!Gender]; N - [mod]- > \\
ADJ; ADJ[upos = ADJ,!Gender]\} \\
without\{N2[upos = ADJ|ADV|INTJ|NOUN|PROPN| \\
VERB|ADP|AUX|CCONJ|DET|NUM|PART|PRON| \\
SCONJ|SYM]; N - [conj]- > N2\} \\
without\{N2[upos = NOUN|PROPN]; N - [appos]- > N2\}
\end{array}
\tag{21.415}$$

(21.415) captures adjectives without gender modifying nouns without gender.

$$\begin{array}{l}
\%31 - AdjN - n - Agree - N \\
pattern\{N[upos = NOUN|PROPN,!Gender]; N - [mod]- > \\
ADJ; ADJ[upos = ADJ, Gender = Fem]\} \\
without\{N2[upos = ADJ|ADV|INTJ|NOUN|PROPN| \\
VERB|ADP|AUX|CCONJ|DET|NUM|PART|PRON| \\
SCONJ|SYM]; N - [conj]- > N2\} \\
without\{N2[upos = NOUN|PROPN]; N - [appos]- > N2\}
\end{array}
\tag{21.416}$$

(21.416) captures adjectives with feminine gender modifying nouns without gender.

$$\begin{array}{l}
\%31 - AdjN - nAgree - Adj \\
pattern\{N[upos = NOUN|PROPN, Gender = Fem]; N - [mod]- > \\
ADJ; ADJ[upos = ADJ,!Gender]\} \\
without\{N2[upos = ADJ|ADV|INTJ|NOUN|PROPN| \\
VERB|ADP|AUX|CCONJ|DET|NUM|PART|PRON| \\
SCONJ|SYM]; N - [conj]- > N2\} \\
without\{N2[upos = NOUN|PROPN]; N - [appos]- > N2\}
\end{array}
\tag{21.417}$$

(21.417) captures adjectives without gender modifying nouns with feminine gender.

$$\begin{array}{l}
 \%31 - AdjN - MascFem \\
 pattern\{N[upos = NOUN|PROPN, Gender = Masc]; N - [mod] - > \\
 ADJ; ADJ[upos = ADJ, Gender = Fem]\} \\
 without\{N2[upos = ADJ|ADV|INTJ|NOUN|PROPN| \\
 VERB|ADP|AUX|CCONJ|DET|NUM|PART|PRON| \\
 SCONJ|SYM]; N - [conj] - > N2\} \\
 without\{N2[upos = NOUN|PROPN]; N - [appos] - > N2\}
 \end{array}
 \tag{21.418}$$

(21.418) captures adjectives with feminine gender modifying nouns with masculine gender.

$$\begin{array}{l}
 \%31 - AdjN - FemMasc \\
 pattern\{N[upos = NOUN|PROPN, Gender = Fem]; N - [mod] - > \\
 ADJ; ADJ[upos = ADJ, Gender = Masc]\} \\
 without\{N2[upos = ADJ|ADV|INTJ|NOUN|PROPN| \\
 VERB|ADP|AUX|CCONJ|DET|NUM|PART|PRON| \\
 SCONJ|SYM]; N - [conj] - > N2\} \\
 without\{N2[upos = NOUN|PROPN]; N - [appos] - > N2\}
 \end{array}
 \tag{21.419}$$

(21.419) captures adjectives with masculine gender modifying nouns with feminine gender.

$$\begin{array}{l}
 \%31 - AdjN - MascMasc \\
 pattern\{N[upos = NOUN|PROPN, Gender = Masc]; N - [mod] - > \\
 ADJ; ADJ[upos = ADJ, Gender = Masc]\} \\
 without\{N2[upos = ADJ|ADV|INTJ|NOUN|PROPN| \\
 VERB|ADP|AUX|CCONJ|DET|NUM|PART|PRON| \\
 SCONJ|SYM]; N - [conj] - > N2\} \\
 without\{N2[upos = NOUN|PROPN]; N - [appos] - > N2\}
 \end{array}
 \tag{21.420}$$

(21.420) captures adjectives with masculine gender modifying nouns with masculine gender.

$$\begin{aligned}
& \%31 - AdjN - FemFem \\
& pattern\{N[upos = NOUN|PROPN, Gender = Fem]; N - [mod] - > \\
& ADJ; ADJ[upos = ADJ, Gender = Fem]\} \\
& without\{N2[upos = ADJ|ADV|INTJ|NOUN|PROPN| \\
& VERB|ADP|AUX|CCONJ|DET|NUM|PART|PRON| \\
& SCONJ|SYM]; N - [conj] - > N2\} \\
& without\{N2[upos = NOUN|PROPN]; N - [appos] - > N2\}
\end{aligned} \tag{21.421}$$

(21.421) captures adjectives with feminine gender modifying nouns with feminine gender.

Subsequently, we want to perform an independent check of the cases of agreement between subject nouns and verbs and, on the other hand, object nouns and verbs, something not foreseen by Greenberg. We believe that this higher level of precision may help to understand better the phenomenon and the distribution of the Universal, which we assume will have higher occurrences in the case of subjects. To do this, it is sufficient to recover the formalizations proposed above, eliminate from the two dependencies activated at the same time the one we are not interested in, and perform this independent search. As an example, we cite what a search for subjects and verbs would be like where there is no agreement in either of the two parts (21.422) and a search for objects and verbs where there is no agreement in either of the two parts (21.423).

$$\begin{aligned}
& \%31 - Subj - Nagree - both \\
& pattern\{V[upos = VERB, !Gender]; V - [subj] - > N; N[upos = \\
& NOUN|PROPN, !Gender]\} \\
& without\{N2[upos = ADJ|ADV|INTJ|NOUN|PROPN| \\
& VERB|ADP|AUX|CCONJ|DET|NUM|PART|PRON| \\
& SCONJ|SYM]; N - [conj] - > N2\} \\
& without\{N2[upos = NOUN|PROPN]; N - [appos] - > N2\}
\end{aligned} \tag{21.422}$$

(21.422) captures subjects without gender and verbs without gender.

$$\begin{aligned}
& \%31 - \textit{Obj} - \textit{Nagree} - \textit{both} \\
& \textit{pattern}\{V[\textit{upos} = \textit{VERB}, !\textit{Gender}]; V - [\textit{comp} : \textit{obj}] - > N; N[\textit{upos} = \\
& \textit{NOUN} | \textit{PROPN}, !\textit{Gender}]\} \\
& \textit{without}\{N2[\textit{upos} = \textit{ADJ} | \textit{ADV} | \textit{INTJ} | \textit{NOUN} | \textit{PROPN} | \\
& \textit{VERB} | \textit{ADP} | \textit{AUX} | \textit{CCONJ} | \textit{DET} | \textit{NUM} | \textit{PART} | \textit{PRON} | \\
& \textit{SCONJ} | \textit{SYM}]; N - [\textit{conj}] - > N2\} \\
& \textit{without}\{N2[\textit{upos} = \textit{NOUN} | \textit{PROPN}]; N - [\textit{appos}] - > N2\}
\end{aligned} \tag{21.423}$$

(21.423) captures objects without gender and verbs without gender.

Other possibilities exist, such as adding the order constraint of the elements cited in the Universal to achieve greater agreement. However, Greenberg makes no mention of this.

21.31.2 Trichotomous Expression Based on Grew-Match Data

Although this Universal is based on specific structures, its formulation is general since its fulfillment can be achieved by means of different variables. We observe a 100% fulfillment of languages of this Universal. In other words, we can establish three different types of languages based on this Universal:

- Languages that do not exhibit any type of gender agreement.
- Languages that exhibit gender agreement in adjectives modifying nouns and no more.
- Languages that exhibit gender agreement in adjectives modifying nouns and subjects or direct objects (or both).

This categorization that we propose, we believe, may be a good starting point for an implicational hierarchy. We propose, therefore, a future study in greater depth of this Universal and of the structures that exhibit gender agreement since, perhaps, a hierarchy greater than the one we propose may emerge.

The most common occurrences we observe are types 1 (no agreement at all) and 3 (both). We must also point out an ambiguity in the formulation of the Universal due to the use of natural language to create these claims. The fact that there is such an agreement between subject and verb, for example, we do not understand it as something that must occur in 100% of the verb tenses but simply mentions the possibility that such a feature exists in some forms or contexts. If

a stricter interpretation is carried out (technically possible), the result would be the same (100%). However, many more languages would be included in type 2 (agreement in adjectives but not subjects/objects).

This information, transferred to the model proposed in this thesis, would appear as in figure 21.68.

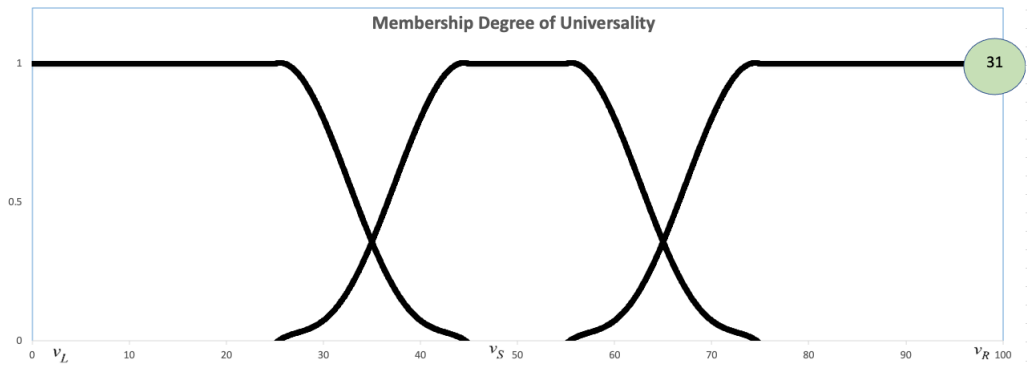


Figure 21.68: Trichotomous Expression for Universal 31

In this case, there are not many aspects to comment on since the undoubted predisposition to be a prototypical High-Universal can be appreciated, and even if exceptions were to appear in the future of languages for which we do not have information today, it seems that such consideration would not change and its high value would be immutable. We can summarise our findings as in table 21.33.

Summary of characteristics of the Universal 31 analysis:
1. Compliance with Greenberg’s Universal is 100%.
2. Greenberg’s Universal is a High-Universal.
3. The fulfillment of the Universal comes, fundamentally, because languages have no gender agreement or have gender agreement between nouns and adjectives and subjects/objects and verbs (in a maximalist interpretation of occurrences).
4. In a stricter interpretation of occurrences, the fulfillment of the Universal would be the same, but there would be more languages ascribed to the other possibility, which is that languages have gender agreement between nouns and adjectives but not between subjects/objects and verbs.

Table 21.33: Summary of Characteristics of the Universal 31 Analysis

21.32 Universal 32

UNIVERSAL 32: *Whenever the verb agrees with a nominal subject or nominal object in gender, it also agrees in number.*

21.32.1 Formalization Proposed for Grew-Match

This Universal is similar to the previously formulated in Universal 31 (21.31). In this case, Greenberg still wants to work with the gender agreement between subjects or objects and the verb. In this case, he is not interested in the agreement with adjectives. On the other hand, he is also interested in the agreement of subjects or objects in number with the verb. Again, there are four possibilities, and the Universal states that only 3 of them are attested in languages: that there is agreement in gender and number, that there is none, or that there is only number agreement. In other words, no languages have only gender agreement between subjects or objects and the verb.

Two different methodologies can do the verification of this Universal. These respond to a verification of the behavior of the language in general and separately (as in the previous Universal) or a verification of the behavior of the particular occurrences of that language with both variables simultaneously. Both seem to be interesting approaches, although we assume that the latter approach may be more problematic since the number of occurrences will be very low in some languages, making it difficult to formulate a Universal. Nevertheless, we formalize both approaches.

In order to carry out the first option, we must reuse the gender-related patterns in subjects and objects from the previous Universal 31 (21.31). We propose again the patterns (21.408) - (21.414). Moreover, again, we do not want to stay only in the generalization in terms of Greenberg, joining subjects and objects, and we also want to analyze separately both dependencies to know their behavior in a more precise way. To do so, we must eliminate the dependence that does not interest us in each case from the above formalizations. In the previous Universal 31, we have shown an example for the subjects and another for the objects of the modification of the case (21.408), using the patterns (21.422) and (21.423). This should apply to all the patterns announced above.

Subsequently, we must capture the behavior of the agreement of number (or not) in the world's different languages, in isolation from gender. Therefore, we propose the patterns (21.424) - (21.427). In these we can clearly distinguish languages that have number from those that do not and, in addition, determine whether there is any agreement between subjects or objects and verbs in

terms of number. We combine, again, the two main labels that languages that have numbers have (as we have done with gender) so as not to show data for number values that are much more restricted to specific languages and, therefore, less Universal. In this sense, we must rely on the hierarchy of number, which details which number values are present in all languages of the world (if they distinguish number), singular and plural. If we were to check, for example, also this Universal through the category of paucal, we would find some languages and occurrences so low that it would not allow us to elaborate any serious claim. Furthermore, according to this category, any language with paucal (or any similar category as dual) has a plural. Therefore, we can check whether a language exhibits number agreement by checking for singulars and plurals without going into other simple addition categories.

The formalizations we propose include subject or object dependency between a verb as head and a noun as dependent. These two elements in each pattern have different features about the number (singular/plural), including all possible combinations between them (with agreement and without). If no occurrences are obtained, the language does not have the feature number in these elements as we previously did with gender in Universal 31. We also add the two restrictions (without) that discard any occurrence formed by the union of two elements by a conjunction or a comma. In the case of the conjunction, we extend to any part of speech since other elements can appear coordinated, not being nouns, as in Spanish (AnCora): “el tema de los rehenes y cuándo tendrá lugar su liberación”, where the head of the second element is a verb. Regarding apposition dependency (comas), less prototypical cases are not taken into account since this dependency must always be established between two nouns in the UD or SUD annotation scheme, as in the example in Spanish (AnCora): “la ministra de Ciencia y Tecnología, Anna Birulés, el alcalde de Barcelona, Joan Clos, la Delegada del Gobierno, Julia García Valdecasas, y una nutrida representación del gobierno catalán”.

We are fully aware that based on feature matches, this agreement check runs the risk of capturing the same matching features but whose motivation is more random and not as motivated as we mentioned in the previous Universal. However, we already stated how this could be solved most objectively. Again, we first propose a general application to subjects and objects, as proposed by Greenberg. Subsequently, we slightly modify the formalizations to independently capture one or the other dependency. For example, we show the separation of the pattern (21.424) into (21.428) for subjects and (21.429) for objects. This same methodology should be applied to the patterns (21.425), (21.426) and (21.427).

$$\begin{array}{l}
 \%32 - \textit{Number} - \textit{AgreeS} \\
 \textit{pattern}\{V[\textit{upos} = \textit{VERB}, \textit{Number} = \textit{Sing}]; V - [\textit{comp} : \textit{obj} | \textit{subj}] - > \\
 N; N[\textit{upos} = \textit{NOUN} | \textit{PROPN}, \textit{Number} = \textit{Sing}]\} \\
 \textit{without}\{N2[\textit{upos} = \textit{ADJ} | \textit{ADV} | \textit{INTJ} | \textit{NOUN} | \textit{PROPN} | \\
 \textit{VERB} | \textit{ADP} | \textit{AUX} | \textit{CCONJ} | \textit{DET} | \textit{NUM} | \textit{PART} | \textit{PRON} | \\
 \textit{SCONJ} | \textit{SYM}]; N - [\textit{conj}] - > N2\} \\
 \textit{without}\{N2[\textit{upos} = \textit{NOUN} | \textit{PROPN}]; N - [\textit{appos}] - > N2\}
 \end{array}
 \tag{21.424}$$

(21.424) captures verbs with singular number and subjects or objects with singular number.

$$\begin{array}{l}
 \%32 - \textit{Number} - \textit{NoAgreeSP} \\
 \textit{pattern}\{V[\textit{upos} = \textit{VERB}, \textit{Number} = \textit{Sing}]; V - [\textit{comp} : \textit{obj} | \textit{subj}] - > \\
 N; N[\textit{upos} = \textit{NOUN} | \textit{PROPN}, \textit{Number} = \textit{Plur}]\} \\
 \textit{without}\{N2[\textit{upos} = \textit{ADJ} | \textit{ADV} | \textit{INTJ} | \textit{NOUN} | \textit{PROPN} | \\
 \textit{VERB} | \textit{ADP} | \textit{AUX} | \textit{CCONJ} | \textit{DET} | \textit{NUM} | \textit{PART} | \textit{PRON} | \\
 \textit{SCONJ} | \textit{SYM}]; N - [\textit{conj}] - > N2\} \\
 \textit{without}\{N2[\textit{upos} = \textit{NOUN} | \textit{PROPN}]; N - [\textit{appos}] - > N2\}
 \end{array}
 \tag{21.425}$$

(21.425) captures verbs with singular number and subjects or objects with plural number.

$$\begin{array}{l}
 \%32 - \textit{Number} - \textit{AgreeP} \\
 \textit{pattern}\{V[\textit{upos} = \textit{VERB}, \textit{Number} = \textit{Plur}]; V - [\textit{comp} : \textit{obj} | \textit{subj}] - > \\
 N; N[\textit{upos} = \textit{NOUN} | \textit{PROPN}, \textit{Number} = \textit{Plur}]\} \\
 \textit{without}\{N2[\textit{upos} = \textit{ADJ} | \textit{ADV} | \textit{INTJ} | \textit{NOUN} | \textit{PROPN} | \\
 \textit{VERB} | \textit{ADP} | \textit{AUX} | \textit{CCONJ} | \textit{DET} | \textit{NUM} | \textit{PART} | \textit{PRON} | \\
 \textit{SCONJ} | \textit{SYM}]; N - [\textit{conj}] - > N2\} \\
 \textit{without}\{N2[\textit{upos} = \textit{NOUN} | \textit{PROPN}]; N - [\textit{appos}] - > N2\}
 \end{array}
 \tag{21.426}$$

(21.426) captures verbs with plural number and subjects or objects with plural number.

$$\begin{array}{l}
\%32 - \text{Number} - \text{NoAgreePS} \\
\text{pattern}\{V[\text{upos} = \text{VERB}, \text{Number} = \text{Plur}]; V - [\text{comp} : \text{obj} | \text{subj}] - > \\
N; N[\text{upos} = \text{NOUN} | \text{PROPN}, \text{Number} = \text{Sing}]\} \\
\text{without}\{N2[\text{upos} = \text{ADJ} | \text{ADV} | \text{INTJ} | \text{NOUN} | \text{PROPN} | \\
\text{VERB} | \text{ADP} | \text{AUX} | \text{CCONJ} | \text{DET} | \text{NUM} | \text{PART} | \text{PRON} | \\
\text{SCONJ} | \text{SYM}]; N - [\text{conj}] - > N2\} \\
\text{without}\{N2[\text{upos} = \text{NOUN} | \text{PROPN}]; N - [\text{appos}] - > N2\}
\end{array}
\tag{21.427}$$

(21.427) captures verbs with plural number and subjects or objects with singular number.

$$\begin{array}{l}
\%32 - \text{NumberS} - \text{AgreeS} \\
\text{pattern}\{V[\text{upos} = \text{VERB}, \text{Number} = \text{Sing}]; V - [\text{subj}] - > \\
N; N[\text{upos} = \text{NOUN} | \text{PROPN}, \text{Number} = \text{Sing}]\} \\
\text{without}\{N2[\text{upos} = \text{ADJ} | \text{ADV} | \text{INTJ} | \text{NOUN} | \text{PROPN} | \\
\text{VERB} | \text{ADP} | \text{AUX} | \text{CCONJ} | \text{DET} | \text{NUM} | \text{PART} | \text{PRON} | \\
\text{SCONJ} | \text{SYM}]; N - [\text{conj}] - > N2\} \\
\text{without}\{N2[\text{upos} = \text{NOUN} | \text{PROPN}]; N - [\text{appos}] - > N2\}
\end{array}
\tag{21.428}$$

(21.428) captures verbs with singular number and subjects with singular number.

$$\begin{array}{l}
\%32 - \text{NumberObj} - \text{AgreeS} \\
\text{pattern}\{V[\text{upos} = \text{VERB}, \text{Number} = \text{Sing}]; V - [\text{comp} : \text{obj}] - > \\
N; N[\text{upos} = \text{NOUN} | \text{PROPN}, \text{Number} = \text{Sing}]\} \\
\text{without}\{N2[\text{upos} = \text{ADJ} | \text{ADV} | \text{INTJ} | \text{NOUN} | \text{PROPN} | \\
\text{VERB} | \text{ADP} | \text{AUX} | \text{CCONJ} | \text{DET} | \text{NUM} | \text{PART} | \text{PRON} | \\
\text{SCONJ} | \text{SYM}]; N - [\text{conj}] - > N2\} \\
\text{without}\{N2[\text{upos} = \text{NOUN} | \text{PROPN}]; N - [\text{appos}] - > N2\}
\end{array}
\tag{21.429}$$

(21.429) captures verbs with singular number and objects with singular number.

On the other hand, we have the possible approach of reviewing occurrences containing a given number feature with a given gender feature. If such an agreement exists, we should always find these two same features in the verb. This is a more direct and reliable way to find agreement of both features simultaneously (and not including separate occurrences). The main problem, as mentioned above, is the low level of occurrences we expect, given the specificity of the search and the low level of verbal gender agreement in the languages. For example, in Russian, verb tenses

that establish a gender distinction in their forms lose the person distinction. If they express a number (plural), they lose gender distinction. There is, therefore, a distribution of features that tend to occur infrequently simultaneously. The patterns we formalize for, in general, subjects and objects is (21.430) - (21.445).

$$\begin{aligned}
 &\%32 - \textit{Gender} - \textit{AgreeFS} \\
 &\textit{pattern}\{V[\textit{upos} = \textit{VERB}, \textit{Gender} = \textit{Fem}, \textit{Number} = \textit{Sing}]; V - \\
 &[\textit{comp} : \textit{obj}|\textit{subj}] - > N; N[\textit{upos} = \textit{NOUN}|\textit{PROPN}, \textit{Gender} = \\
 &\textit{Fem}, \textit{Number} = \textit{Sing}]\} \\
 &\textit{without}\{N2[\textit{upos} = \textit{ADJ}|\textit{ADV}|\textit{INTJ}|\textit{NOUN}|\textit{PROPN}| \\
 &\textit{VERB}|\textit{ADP}|\textit{AUX}|\textit{CCONJ}|\textit{DET}|\textit{NUM}|\textit{PART}|\textit{PRON}| \\
 &\textit{SCONJ}|\textit{SYM}]; N - [\textit{conj}] - > N2\} \\
 &\textit{without}\{N2[\textit{upos} = \textit{NOUN}|\textit{PROPN}]; N - [\textit{appos}] - > N2\}
 \end{aligned}
 \tag{21.430}$$

(21.430) captures verbs with feminine gender and singular number and objects or subjects with feminine gender and singular number.

$$\begin{aligned}
 &\%32 - \textit{Gender} - \textit{AgreeFP} \\
 &\textit{pattern}\{V[\textit{upos} = \textit{VERB}, \textit{Gender} = \textit{Fem}, \textit{Number} = \textit{Plur}]; V - \\
 &[\textit{comp} : \textit{obj}|\textit{subj}] - > N; N[\textit{upos} = \textit{NOUN}|\textit{PROPN}, \textit{Gender} = \\
 &\textit{Fem}, \textit{Number} = \textit{Plur}]\} \\
 &\textit{without}\{N2[\textit{upos} = \textit{ADJ}|\textit{ADV}|\textit{INTJ}|\textit{NOUN}|\textit{PROPN}| \\
 &\textit{VERB}|\textit{ADP}|\textit{AUX}|\textit{CCONJ}|\textit{DET}|\textit{NUM}|\textit{PART}|\textit{PRON}| \\
 &\textit{SCONJ}|\textit{SYM}]; N - [\textit{conj}] - > N2\} \\
 &\textit{without}\{N2[\textit{upos} = \textit{NOUN}|\textit{PROPN}]; N - [\textit{appos}] - > N2\}
 \end{aligned}
 \tag{21.431}$$

(21.431) captures verbs with feminine gender and plural number and objects or subjects with feminine gender and plural number.

$$\begin{aligned}
&\%32 - \textit{Gender} - \textit{nAgreeFP} \\
&\textit{pattern}\{V[\textit{upos} = \textit{VERB}, \textit{Gender} = \textit{Fem}, \textit{Number} = \textit{Plur}]; V - \\
&[\textit{comp} : \textit{obj}|\textit{subj}] - > N; N[\textit{upos} = \textit{NOUN}|\textit{PROPN}, \textit{Gender} = \\
&\textit{Fem}, \textit{Number} = \textit{Sing}]\} \\
&\textit{without}\{N2[\textit{upos} = \textit{ADJ}|\textit{ADV}|\textit{INTJ}|\textit{NOUN}|\textit{PROPN}| \\
&\textit{VERB}|\textit{ADP}|\textit{AUX}|\textit{CCONJ}|\textit{DET}|\textit{NUM}|\textit{PART}|\textit{PRON}| \\
&\textit{SCONJ}|\textit{SYM}]; N - [\textit{conj}] - > N2\} \\
&\textit{without}\{N2[\textit{upos} = \textit{NOUN}|\textit{PROPN}]; N - [\textit{appos}] - > N2\}
\end{aligned} \tag{21.432}$$

(21.432) captures verbs with feminine gender and plural number and objects or subjects with feminine gender and singular number.

$$\begin{aligned}
&\%32 - \textit{Gender} - \textit{nAgreeFS} \\
&\textit{pattern}\{V[\textit{upos} = \textit{VERB}, \textit{Gender} = \textit{Fem}, \textit{Number} = \textit{Sing}]; V - \\
&[\textit{comp} : \textit{obj}|\textit{subj}] - > N; N[\textit{upos} = \textit{NOUN}|\textit{PROPN}, \textit{Gender} = \\
&\textit{Fem}, \textit{Number} = \textit{Plur}]\} \\
&\textit{without}\{N2[\textit{upos} = \textit{ADJ}|\textit{ADV}|\textit{INTJ}|\textit{NOUN}|\textit{PROPN}| \\
&\textit{VERB}|\textit{ADP}|\textit{AUX}|\textit{CCONJ}|\textit{DET}|\textit{NUM}|\textit{PART}|\textit{PRON}| \\
&\textit{SCONJ}|\textit{SYM}]; N - [\textit{conj}] - > N2\} \\
&\textit{without}\{N2[\textit{upos} = \textit{NOUN}|\textit{PROPN}]; N - [\textit{appos}] - > N2\}
\end{aligned} \tag{21.433}$$

(21.433) captures verbs with feminine gender and singular number and objects or subjects with feminine gender and plural number.

$$\begin{aligned}
&\$32 - \textit{Gender} - \textit{AgreeMS} \\
&\textit{pattern}\{V[\textit{upos} = \textit{VERB}, \textit{Gender} = \textit{Masc}, \textit{Number} = \textit{Sing}]; V - \\
&[\textit{comp} : \textit{obj}|\textit{subj}] - > N; N[\textit{upos} = \textit{NOUN}|\textit{PROPN}, \textit{Gender} = \\
&\textit{Masc}, \textit{Number} = \textit{Sing}]\} \\
&\textit{without}\{N2[\textit{upos} = \textit{ADJ}|\textit{ADV}|\textit{INTJ}|\textit{NOUN}|\textit{PROPN}| \\
&\textit{VERB}|\textit{ADP}|\textit{AUX}|\textit{CCONJ}|\textit{DET}|\textit{NUM}|\textit{PART}|\textit{PRON}| \\
&\textit{SCONJ}|\textit{SYM}]; N - [\textit{conj}] - > N2\} \\
&\textit{without}\{N2[\textit{upos} = \textit{NOUN}|\textit{PROPN}]; N - [\textit{appos}] - > N2\}
\end{aligned} \tag{21.434}$$

(21.434) captures verbs with masculine gender and singular number and objects or subjects with masculine gender and singular number.

$$\begin{aligned}
 &\%32 - \textit{Gender} - \textit{AgreeMP} \\
 &\textit{pattern}\{V[\textit{upos} = \textit{VERB}, \textit{Gender} = \textit{Masc}, \textit{Number} = \textit{Plur}]; V - \\
 &[\textit{comp} : \textit{obj}|\textit{subj}] - > N; N[\textit{upos} = \textit{NOUN}|\textit{PROPN}, \textit{Gender} = \\
 &\textit{Masc}, \textit{Number} = \textit{Plur}]\} \\
 &\textit{without}\{N2[\textit{upos} = \textit{ADJ}|\textit{ADV}|\textit{INTJ}|\textit{NOUN}|\textit{PROPN}| \\
 &\textit{VERB}|\textit{ADP}|\textit{AUX}|\textit{CCONJ}|\textit{DET}|\textit{NUM}|\textit{PART}|\textit{PRON}| \\
 &\textit{SCONJ}|\textit{SYM}]; N - [\textit{conj}] - > N2\} \\
 &\textit{without}\{N2[\textit{upos} = \textit{NOUN}|\textit{PROPN}]; N - [\textit{appos}] - > N2\}
 \end{aligned}
 \tag{21.435}$$

(21.435) captures verbs with masculine gender and plural number and objects or subjects with masculine gender and plural number.

$$\begin{aligned}
 &\%32 - \textit{Gender} - \textit{nAgreeMP} \\
 &\textit{pattern}\{V[\textit{upos} = \textit{VERB}, \textit{Gender} = \textit{Masc}, \textit{Number} = \textit{Plur}]; V - \\
 &[\textit{comp} : \textit{obj}|\textit{subj}] - > N; N[\textit{upos} = \textit{NOUN}|\textit{PROPN}, \textit{Gender} = \\
 &\textit{Masc}, \textit{Number} = \textit{Sing}]\} \\
 &\textit{without}\{N2[\textit{upos} = \textit{ADJ}|\textit{ADV}|\textit{INTJ}|\textit{NOUN}|\textit{PROPN}| \\
 &\textit{VERB}|\textit{ADP}|\textit{AUX}|\textit{CCONJ}|\textit{DET}|\textit{NUM}|\textit{PART}|\textit{PRON}| \\
 &\textit{SCONJ}|\textit{SYM}]; N - [\textit{conj}] - > N2\} \\
 &\textit{without}\{N2[\textit{upos} = \textit{NOUN}|\textit{PROPN}]; N - [\textit{appos}] - > N2\}
 \end{aligned}
 \tag{21.436}$$

(21.436) captures verbs with masculine gender and plural number and objects or subjects with masculine gender and singular number.

$$\begin{aligned}
 &\%32 - \textit{Gender} - \textit{nAgreeFS} \\
 &\textit{pattern}\{V[\textit{upos} = \textit{VERB}, \textit{Gender} = \textit{Masc}, \textit{Number} = \textit{Sing}]; V - \\
 &[\textit{comp} : \textit{obj}|\textit{subj}] - > N; N[\textit{upos} = \textit{NOUN}|\textit{PROPN}, \textit{Gender} = \\
 &\textit{Masc}, \textit{Number} = \textit{Plur}]\} \\
 &\textit{without}\{N2[\textit{upos} = \textit{ADJ}|\textit{ADV}|\textit{INTJ}|\textit{NOUN}|\textit{PROPN}| \\
 &\textit{VERB}|\textit{ADP}|\textit{AUX}|\textit{CCONJ}|\textit{DET}|\textit{NUM}|\textit{PART}|\textit{PRON}| \\
 &\textit{SCONJ}|\textit{SYM}]; N - [\textit{conj}] - > N2\} \\
 &\textit{without}\{N2[\textit{upos} = \textit{NOUN}|\textit{PROPN}]; N - [\textit{appos}] - > N2\}
 \end{aligned}
 \tag{21.437}$$

(21.433) captures verbs with masculine gender and singular number and objects or subjects with masculine gender and plural number.

$$\begin{aligned}
& \%32 - \textit{Gender} - n\textit{AgreeFM} - sg \\
& \textit{pattern}\{V[\textit{upos} = \textit{VERB}, \textit{Gender} = \textit{Fem}, \textit{Number} = \textit{Sing}]; V - \\
& [\textit{comp} : \textit{obj}|\textit{subj}] - > N; N[\textit{upos} = \textit{NOUN}|\textit{PROPN}, \textit{Gender} = \\
& \textit{Masc}, \textit{Number} = \textit{Sing}]\} \\
& \textit{without}\{N2[\textit{upos} = \textit{ADJ}|\textit{ADV}|\textit{INTJ}|\textit{NOUN}|\textit{PROPN}| \\
& \textit{VERB}|\textit{ADP}|\textit{AUX}|\textit{CCONJ}|\textit{DET}|\textit{NUM}|\textit{PART}|\textit{PRON}| \\
& \textit{SCONJ}|\textit{SYM}]; N - [\textit{conj}] - > N2\} \\
& \textit{without}\{N2[\textit{upos} = \textit{NOUN}|\textit{PROPN}]; N - [\textit{appos}] - > N2\}
\end{aligned} \tag{21.438}$$

(21.438) captures verbs with feminine gender and singular number and objects or subjects with masculine gender and singular number.

$$\begin{aligned}
& \%32 - \textit{Gender} - n\textit{AgreeMF} - sg \\
& \textit{pattern}\{V[\textit{upos} = \textit{VERB}, \textit{Gender} = \textit{Masc}, \textit{Number} = \textit{Sing}]; V - \\
& [\textit{comp} : \textit{obj}|\textit{subj}] - > N; N[\textit{upos} = \textit{NOUN}|\textit{PROPN}, \textit{Gender} = \\
& \textit{Fem}, \textit{Number} = \textit{Sing}]\} \\
& \textit{without}\{N2[\textit{upos} = \textit{ADJ}|\textit{ADV}|\textit{INTJ}|\textit{NOUN}|\textit{PROPN}| \\
& \textit{VERB}|\textit{ADP}|\textit{AUX}|\textit{CCONJ}|\textit{DET}|\textit{NUM}|\textit{PART}|\textit{PRON}| \\
& \textit{SCONJ}|\textit{SYM}]; N - [\textit{conj}] - > N2\} \\
& \textit{without}\{N2[\textit{upos} = \textit{NOUN}|\textit{PROPN}]; N - [\textit{appos}] - > N2\}
\end{aligned} \tag{21.439}$$

(21.439) captures verbs with masculine gender and singular number and objects or subjects with feminine gender and singular number.

$$\begin{aligned}
& \%32 - \textit{Gender} - n\textit{AgreeFM} - pl \\
& \textit{pattern}\{V[\textit{upos} = \textit{VERB}, \textit{Gender} = \textit{Fem}, \textit{Number} = \\
& \textit{Plur}]; V - [\textit{comp} : \textit{obj}|\textit{subj}] - > N; N[\textit{upos} = \\
& \textit{NOUN}|\textit{PROPN}, \textit{Gender} = \textit{Masc}, \textit{Number} = \\
& \textit{Plur}]\} \textit{without}\{N2[\textit{upos} = \textit{ADJ}|\textit{ADV}|\textit{INTJ}|\textit{NOUN}|\textit{PROPN}| \\
& \textit{VERB}|\textit{ADP}|\textit{AUX}|\textit{CCONJ}|\textit{DET}|\textit{NUM}|\textit{PART}|\textit{PRON}| \\
& \textit{SCONJ}|\textit{SYM}]; N - [\textit{conj}] - > N2\} \\
& \textit{without}\{N2[\textit{upos} = \textit{NOUN}|\textit{PROPN}]; N - [\textit{appos}] - > N2\}
\end{aligned} \tag{21.440}$$

(21.440) captures verbs with feminine gender and plural number and objects or subjects with masculine gender and plural number.

$$\begin{aligned}
& \%32 - \textit{Gender} - n\textit{AgreeMF} - \textit{pl} \\
& \textit{pattern}\{V[\textit{upos} = \textit{VERB}, \textit{Gender} = \textit{Masc}, \textit{Number} = \textit{Plur}]; V - \\
& [\textit{comp} : \textit{obj}|\textit{subj}] - > N; N[\textit{upos} = \textit{NOUN}|\textit{PROPN}, \textit{Gender} = \\
& \textit{Fem}, \textit{Number} = \textit{Plur}]\} \\
& \textit{without}\{N2[\textit{upos} = \textit{ADJ}|\textit{ADV}|\textit{INTJ}|\textit{NOUN}|\textit{PROPN}| \\
& \textit{VERB}|\textit{ADP}|\textit{AUX}|\textit{CCONJ}|\textit{DET}|\textit{NUM}|\textit{PART}|\textit{PRON}| \\
& \textit{SCONJ}|\textit{SYM}]; N - [\textit{conj}] - > N2\} \\
& \textit{without}\{N2[\textit{upos} = \textit{NOUN}|\textit{PROPN}]; N - [\textit{appos}] - > N2\}
\end{aligned}
\tag{21.441}$$

(21.441) captures verbs with masculine gender and plural number and objects or subjects with feminine gender and plural number.

$$\begin{aligned}
& \%32 - \textit{Gender} - n\textit{AgreeFM} - \textit{PS} \\
& \textit{pattern}\{V[\textit{upos} = \textit{VERB}, \textit{Gender} = \textit{Fem}, \textit{Number} = \textit{Plur}]; V - \\
& [\textit{comp} : \textit{obj}|\textit{subj}] - > N; N[\textit{upos} = \textit{NOUN}|\textit{PROPN}, \textit{Gender} = \\
& \textit{Masc}, \textit{Number} = \textit{Sing}]\} \\
& \textit{without}\{N2[\textit{upos} = \textit{ADJ}|\textit{ADV}|\textit{INTJ}|\textit{NOUN}|\textit{PROPN}| \\
& \textit{VERB}|\textit{ADP}|\textit{AUX}|\textit{CCONJ}|\textit{DET}|\textit{NUM}|\textit{PART}|\textit{PRON}| \\
& \textit{SCONJ}|\textit{SYM}]; N - [\textit{conj}] - > N2\} \\
& \textit{without}\{N2[\textit{upos} = \textit{NOUN}|\textit{PROPN}]; N - [\textit{appos}] - > N2\}
\end{aligned}
\tag{21.442}$$

(21.442) captures verbs with feminine gender and plural number and objects or subjects with masculine gender and singular number.

$$\begin{aligned}
& \%32 - \textit{Gender} - n\textit{AgreeMF} - \textit{PS} \\
& \textit{pattern}\{V[\textit{upos} = \textit{VERB}, \textit{Gender} = \textit{Masc}, \textit{Number} = \textit{Plur}]; V - \\
& [\textit{comp} : \textit{obj}|\textit{subj}] - > N; N[\textit{upos} = \textit{NOUN}|\textit{PROPN}, \textit{Gender} = \\
& \textit{Fem}, \textit{Number} = \textit{Sing}]\} \\
& \textit{without}\{N2[\textit{upos} = \textit{ADJ}|\textit{ADV}|\textit{INTJ}|\textit{NOUN}|\textit{PROPN}| \\
& \textit{VERB}|\textit{ADP}|\textit{AUX}|\textit{CCONJ}|\textit{DET}|\textit{NUM}|\textit{PART}|\textit{PRON}| \\
& \textit{SCONJ}|\textit{SYM}]; N - [\textit{conj}] - > N2\} \\
& \textit{without}\{N2[\textit{upos} = \textit{NOUN}|\textit{PROPN}]; N - [\textit{appos}] - > N2\}
\end{aligned}
\tag{21.443}$$

(21.443) captures verbs with masculine gender and plural number and objects or subjects with feminine gender and singular number.

$$\begin{aligned}
& \%32 - \textit{Gender} - n\textit{AgreeFM} - SP \\
& \textit{pattern}\{V[\textit{upos} = \textit{VERB}, \textit{Gender} = \textit{Fem}, \textit{Number} = \textit{Sing}]; V - \\
& [\textit{comp} : \textit{obj}|\textit{subj}] - > N; N[\textit{upos} = \textit{NOUN}|\textit{PROPN}, \textit{Gender} = \\
& \textit{Masc}, \textit{Number} = \textit{Plur}]\} \\
& \textit{without}\{N2[\textit{upos} = \textit{ADJ}|\textit{ADV}|\textit{INTJ}|\textit{NOUN}|\textit{PROPN}| \\
& \textit{VERB}|\textit{ADP}|\textit{AUX}|\textit{CCONJ}|\textit{DET}|\textit{NUM}|\textit{PART}|\textit{PRON}| \\
& \textit{SCONJ}|\textit{SYM}]; N - [\textit{conj}] - > N2\} \\
& \textit{without}\{N2[\textit{upos} = \textit{NOUN}|\textit{PROPN}]; N - [\textit{appos}] - > N2\}
\end{aligned} \tag{21.444}$$

(21.444) captures verbs with feminine gender and singular number and objects or subjects with masculine gender and plural number.

$$\begin{aligned}
& \%32 - \textit{Gender} - n\textit{AgreeMF} - SP \\
& \textit{pattern}\{V[\textit{upos} = \textit{VERB}, \textit{Gender} = \textit{Masc}, \textit{Number} = \textit{Sing}]; V - \\
& [\textit{comp} : \textit{obj}|\textit{subj}] - > N; N[\textit{upos} = \textit{NOUN}|\textit{PROPN}, \textit{Gender} = \\
& \textit{Fem}, \textit{Number} = \textit{Plur}]\} \\
& \textit{without}\{N2[\textit{upos} = \textit{ADJ}|\textit{ADV}|\textit{INTJ}|\textit{NOUN}|\textit{PROPN}| \\
& \textit{VERB}|\textit{ADP}|\textit{AUX}|\textit{CCONJ}|\textit{DET}|\textit{NUM}|\textit{PART}|\textit{PRON}| \\
& \textit{SCONJ}|\textit{SYM}]; N - [\textit{conj}] - > N2\} \\
& \textit{without}\{N2[\textit{upos} = \textit{NOUN}|\textit{PROPN}]; N - [\textit{appos}] - > N2\}
\end{aligned} \tag{21.445}$$

(21.445) captures verbs with masculine gender and singular number and objects or subjects with feminine gender and plural number.

Again, in this case, we propose creating new patterns based on the ones described above, applied separately to subjects and direct objects. As an example, we convert pattern (21.430) into patterns (21.446) (subject) and (21.447) (object). This should apply equally to the rest.

$$\begin{aligned}
 &\%32 - S\text{Gender} - AgreeFS \\
 &pattern\{V[upos = VERB, Gender = Fem, Number = Sing]; V - \\
 &[subj] - > N; N[upos = NOUN|PROPN, Gender = Fem, Number = \\
 &Sing]\} \\
 &without\{N2[upos = ADJ|ADV|INTJ|NOUN|PROPN| \\
 &VERB|ADP|AUX|CCONJ|DET|NUM|PART|PRON| \\
 &SCONJ|SYM]; N - [conj] - > N2\} \\
 &without\{N2[upos = NOUN|PROPN]; N - [appos] - > N2\}
 \end{aligned}
 \tag{21.446}$$

(21.446) captures verbs with feminine gender and singular number and subjects with feminine gender and singular number.

$$\begin{aligned}
 &\%32 - O\text{Gender} - AgreeFS \\
 &pattern\{V[upos = VERB, Gender = Fem, Number = Sing]; V - \\
 &[comp : obj] - > N; N[upos = NOUN|PROPN, Gender = \\
 &Fem, Number = Sing]\} \\
 &without\{N2[upos = ADJ|ADV|INTJ|NOUN|PROPN| \\
 &VERB|ADP|AUX|CCONJ|DET|NUM|PART|PRON| \\
 &SCONJ|SYM]; N - [conj] - > N2\} \\
 &without\{N2[upos = NOUN|PROPN]; N - [appos] - > N2\}
 \end{aligned}
 \tag{21.447}$$

(21.447) captures verbs with feminine gender and singular number and objects with feminine gender and singular number.

21.32.2 Trichotomous Expression Based on Grew-Match Data

This Universal, also based on agreements, is again based on fulfilling a condition to be true. In this case, on the other hand, what is observed is the primacy of the number agreement over the gender agreement. This gives rise, once again, to a possible classification of linguistic types into:

- Languages that exhibit neither of the two agreements.
- Languages that exhibit number agreement but not gender agreement in either the subject or the object.
- Languages that exhibit both number agreement and gender agreement in the subject or object.

In this case, most of the completions are produced by subject dependency. We would also like to point out that this is a clear example of a Universal that can be easily checked in a grammar-based approach or by means of tests. However, given that such approaches have already been done throughout history, we believe that the data expressed here, although they may contain some difficulties during processing, offer a very interesting counterpoint. This is especially true when dealing with non-canonical but existing structures in languages (such as the gender agreement between subject and verb in Romance languages), which are not usually included in other types of approaches.

One of the possibilities that also allows a quantitative approach (which is not developed in this thesis because it is not its primary objective, but it is postulated as a future objective) is to contrast the quantitative proportions of the occurrences of the different types. That is, to check whether, for example, in the case of languages exhibiting both number agreement and gender agreement, the proportion of occurrences is always higher in the case of number. This distinct check would help to reinforce the idea of a higher relevance of number agreement compared to gender agreement from another metric.

The tracking of this Universal is in a 91% or, in Fuzzy Logic terms, as in figure 21.69.

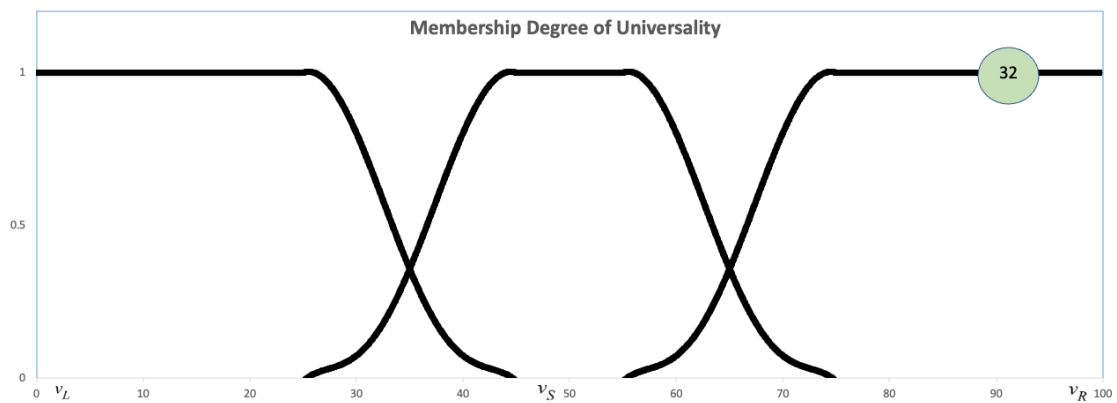


Figure 21.69: Trichotomous Expression for Universal 32

Again, we are dealing with a prototypical High-Universal. Therefore, this is another rule that allows us to understand better, classify, and predict the behavior of the world's languages. We can summarise our findings as in table 21.34.

Summary of characteristics of the Universal 32 analysis: 1. There is a primacy of number agreement over gender agreement. 2. It is easier to check Universals of this kind through grammars or through introspection. However, it is possible to capture non-canonical structures in unfamiliar languages that might go unnoticed. 3. The fulfillment of Greenberg’s original Universal is 91%. 4. The original Greenberg Universal is a High-Universal.

Table 21.34: Summary of Characteristics of the Universal 32 Analysis

21.33 Universal 33

UNIVERSAL 33: *When number agreement between the noun and verb is suspended and the rules is based on order, the case is always one in which the verb precedes and the verb is singular.*

21.33.1 Formalization Proposed for Grew-Match

This Universal is challenging to analyze, mainly because of the wording proposed by Greenberg. First of all, although he does not make it explicit, the agreement between nouns and verbs refers to the two previously formulated dependencies applicable to subjects and objects. In this case, the Universal applies to languages in which there is no number agreement between subjects/objects and verbs. Let us recall that this can manifest itself in two different ways: first, it may be that in a language, this feature does not exist (and therefore does not appear in the information of the word any number feature). There is also the case that such information exists in the language but that it should not coincide between both parts of speech; that is, they disagree. It is tough to determine when the agreement between the noun and the verb is suspended, and the rule is based on order. It is challenging to determine that there is an order restriction in specific cases if one is not an expert in the language.

Therefore, we try to collect all the possibilities we have. In all possibilities, we must combine the absence of the feature, the singular value, and the plural value between verbs and nouns (subj and obj). In addition, as we have done so far, we also check subj and obj separately. In this case,

since it is mentioned in the Universal, we add an order constraint in which the verb can precede the noun (which Greenberg indicates occurs) and in which the noun can precede the verb (which Greenberg indicates does not occur). This means that we double the patterns created.

Regarding the obligation of being singular, we may find inconsistencies in the tagging; in some treebanks, we may find the singular tag, and in others, we may not find a tag because there is no number value since both may correspond to the same situation. This problem already appears in Greenberg since his research mentions “singular” as a synonym of the unmarked version of the word. Therefore, the annotator of each language must consider whether the unmarked form we are observing corresponds to a verb with the genuine singular number feature or, on the contrary, simply has no feature at all. On the one hand, we believe that experts in each language can easily disambiguate these cases when annotating. On the other hand, it would not be strange to find some inconsistencies.

This makes it necessary to mark some patterns as possible and others as impossible, according to Greenberg. For example, in cases where the noun has no number feature, the verb should not have a number feature either, or it should have the singular feature and be placed before the affected noun. In the cases in which the noun has a plural number feature, the verb may have no label or may have the singular label if it is placed before the noun, or it may have the plural label (there would be agreement, and it would not count for this Universal language), with a free position. If the noun has a singular number feature, the verb may be in the singular or not have an agreement tag. In the case of not having a label, it must always appear before it. In the case of having the singular tag, we cannot determine whether it appears with that value because it has agreement or describes Greenberg in its Universal, so we cannot establish an order restriction.

In summary, we must apply this Universal to all languages, which, in previous Universals, we have certified as having some nominal element (subject or object) without number agreement with the verb. The feature of the noun can be any in those cases (most probably, it does not have a feature either), but the verb must always appear without value or with singular value (we will also see the distribution of these two options among the results) and appear before the noun.

In order to elaborate the formalizations, we create patterns with the two elements mentioned above, verbs and nouns (common or proper nouns). We establish a dependency relationship in which the verb is the head and the noun is the dependent. This relationship can be subject or object. In addition, in one pattern, we ask that the verb is in the first position and, in another, that the noun is in the first position. We continue adding two negative constraints (without)

that discard more atypical cases with coordination and similar cases that can alter the captured image. In each pattern, we modify the distribution of the number of features between verbs and nouns, combining the previously mentioned possibilities: no feature, singular, and plural. This corresponds to (21.448) - (21.465).

$$\begin{array}{l}
 \%33 - VN - general - nofeatVN - Vbef \\
 pattern\{V[upos = VERB, !Number]; N[upos = \\
 NOUN|PROPN, !Number]; V - [comp : obj|subj] - > N; V << N\} \\
 without\{N2[upos = ADJ|ADV|INTJ|NOUN|PROPN| \\
 VERB|ADP|AUX|CCONJ|DET|NUM|PART|PRON| \\
 SCONJ|SYM]; N - [conj] - > N2\} \\
 without\{N2[upos = NOUN|PROPN]; N - [appos] - > N2\}
 \end{array} \tag{21.448}$$

(21.448) captures verbs without number before the nominal subjects or objects without number.

$$\begin{array}{l}
 \%33 - VN - general - nofeatVN - Vaft \\
 pattern\{V[upos = VERB, !Number]; N[upos = \\
 NOUN|PROPN, !Number]; V - [comp : obj|subj] - > N; V >> N\} \\
 without\{N2[upos = ADJ|ADV|INTJ|NOUN|PROPN| \\
 VERB|ADP|AUX|CCONJ|DET|NUM|PART|PRON| \\
 SCONJ|SYM]; N - [conj] - > N2\} \\
 without\{N2[upos = NOUN|PROPN]; N - [appos] - > N2\}
 \end{array} \tag{21.449}$$

(21.449) captures verbs without number after the nominal subjects or objects without number.

$$\begin{array}{l}
 \%33 - VN - general - nofeatV - Vbef - Nsing \\
 pattern\{V[upos = VERB, !Number]; N[upos = \\
 NOUN|PROPN, Number = Sing]; V - [comp : obj|subj] - > \\
 N; V << N\} \\
 without\{N2[upos = ADJ|ADV|INTJ|NOUN|PROPN| \\
 VERB|ADP|AUX|CCONJ|DET|NUM|PART|PRON| \\
 SCONJ|SYM]; N - [conj] - > N2\} \\
 without\{N2[upos = NOUN|PROPN]; N - [appos] - > N2\}
 \end{array} \tag{21.450}$$

(21.450) captures verbs without number before the nominal subjects or objects with singular number.

$$\begin{array}{l}
\%33 - VN - general - nofeatV - Vaft - Nsing \\
pattern\{V[upos = VERB, !Number]; N[upos = \\
NOUN|PROPN, Number = Sing]; V - [comp : obj|subj] - > \\
N; V >> N\} \\
without\{N2[upos = ADJ|ADV|INTJ|NOUN|PROPN| \\
VERB|ADP|AUX|CCONJ|DET|NUM|PART|PRON| \\
SCONJ|SYM]; N - [conj] - > N2\} \\
without\{N2[upos = NOUN|PROPN]; N - [appos] - > N2\}
\end{array}
\tag{21.451}$$

(21.451) captures verbs without number after the nominal subjects or objects with singular number.

$$\begin{array}{l}
\%33 - VN - general - nofeatV - Vbef - Nplur \\
pattern\{V[upos = VERB, !Number]; N[upos = \\
NOUN|PROPN, Number = Plur]; V - [comp : obj|subj] - > \\
N; V << N\} \\
without\{N2[upos = ADJ|ADV|INTJ|NOUN|PROPN| \\
VERB|ADP|AUX|CCONJ|DET|NUM|PART|PRON| \\
SCONJ|SYM]; N - [conj] - > N2\} \\
without\{N2[upos = NOUN|PROPN]; N - [appos] - > N2\}
\end{array}
\tag{21.452}$$

(21.452) captures verbs without number before the nominal subjects or objects with plural number.

$$\begin{array}{l}
\%33 - VN - general - nofeatV - Vaft - Nplur \\
pattern\{V[upos = VERB, !Number]; N[upos = \\
NOUN|PROPN, Number = Plur]; V - [comp : obj|subj] - > \\
N; V >> N\} \\
without\{N2[upos = ADJ|ADV|INTJ|NOUN|PROPN| \\
VERB|ADP|AUX|CCONJ|DET|NUM|PART|PRON| \\
SCONJ|SYM]; N - [conj] - > N2\} \\
without\{N2[upos = NOUN|PROPN]; N - [appos] - > N2\}
\end{array}
\tag{21.453}$$

(21.453) captures verbs without number after the nominal subjects or objects with plural number.

$$\begin{aligned}
&\%33 - VN - general - nofeatN - Vbef - Vsing \\
&pattern\{V[upos = VERB, Number = Sing]; N[upos = \\
&NOUN|PROPN, !Number]; V - [comp : obj|subj] - > N; V << N\} \\
&without\{N2[upos = ADJ|ADV|INTJ|NOUN|PROPN| \\
&VERB|ADP|AUX|CCONJ|DET|NUM|PART|PRON| \\
&SCONJ|SYM]; N - [conj] - > N2\} \\
&without\{N2[upos = NOUN|PROPN]; N - [appos] - > N2\}
\end{aligned}
\tag{21.454}$$

(21.454) captures verbs with singular number before the nominal subjects or objects without number.

$$\begin{aligned}
&\%33 - VN - general - nofeatN - Vaft - Vsing \\
&pattern\{V[upos = VERB, Number = Sing]; N[upos = \\
&NOUN|PROPN, !Number]; V - [comp : obj|subj] - > N; V >> N\} \\
&without\{N2[upos = ADJ|ADV|INTJ|NOUN|PROPN| \\
&VERB|ADP|AUX|CCONJ|DET|NUM|PART|PRON| \\
&SCONJ|SYM]; N - [conj] - > N2\} \\
&without\{N2[upos = NOUN|PROPN]; N - [appos] - > N2\}
\end{aligned}
\tag{21.455}$$

(21.455) captures verbs with singular number after the nominal subjects or objects without number.

$$\begin{aligned}
&\%33 - VN - general - nofeatN - Vbef - Vplur \\
&pattern\{V[upos = VERB, Number = Plur]; N[upos = \\
&NOUN|PROPN, !Number]; V - [comp : obj|subj] - > N; V << N\} \\
&without\{N2[upos = ADJ|ADV|INTJ|NOUN|PROPN| \\
&VERB|ADP|AUX|CCONJ|DET|NUM|PART|PRON| \\
&SCONJ|SYM]; N - [conj] - > N2\} \\
&without\{N2[upos = NOUN|PROPN]; N - [appos] - > N2\}
\end{aligned}
\tag{21.456}$$

(21.456) captures verbs with plural number before the nominal subjects or objects without number.

$$\begin{aligned}
&\%33 - VN - general - nofeatN - Vaft - Vplur \\
&pattern\{V[upos = VERB, Number = Plur]; N[upos = \\
&NOUN|PROPN, !Number]; V - [comp : obj|subj] - > N; V >> N\} \\
&without\{N2[upos = ADJ|ADV|INTJ|NOUN|PROPN| \\
&VERB|ADP|AUX|CCONJ|DET|NUM|PART|PRON| \\
&SCONJ|SYM]; N - [conj] - > N2\} \\
&without\{N2[upos = NOUN|PROPN]; N - [appos] - > N2\}
\end{aligned}
\tag{21.457}$$

(21.457) captures verbs with plural number after the nominal subjects or objects without number.

$$\begin{aligned}
&\%33 - VN - general - VplNpl - Vbef \\
&pattern\{V[upos = VERB, Number = Plur]; N[upos = \\
&NOUN|PROPN, Number = Plur]; V - [comp : obj|subj] - > \\
&N; V << N\} \\
&without\{N2[upos = ADJ|ADV|INTJ|NOUN|PROPN| \\
&VERB|ADP|AUX|CCONJ|DET|NUM|PART|PRON| \\
&SCONJ|SYM]; N - [conj] - > N2\} \\
&without\{N2[upos = NOUN|PROPN]; N - [appos] - > N2\}
\end{aligned}
\tag{21.458}$$

(21.458) captures verbs with plural number before the nominal subjects or objects.

$$\begin{aligned}
&\%33 - VN - general - VplNpl - Vaft \\
&pattern\{V[upos = VERB, Number = Plur]; N[upos = \\
&NOUN|PROPN, Number = Plur]; V - [comp : obj|subj] - > \\
&N; V >> N\} \\
&without\{N2[upos = ADJ|ADV|INTJ|NOUN|PROPN| \\
&VERB|ADP|AUX|CCONJ|DET|NUM|PART|PRON| \\
&SCONJ|SYM]; N - [conj] - > N2\} \\
&without\{N2[upos = NOUN|PROPN]; N - [appos] - > N2\}
\end{aligned}
\tag{21.459}$$

(21.459) captures verbs with plural number after the nominal subjects or objects with plural number.

$$\begin{aligned}
&\%33 - VN - general - VsinNsin - Vbef \\
&pattern\{V[upos = VERB, Number = Sing]; N[upos = \\
&NOUN|PROPN, Number = Sing]; V - [comp : obj|subj] - > \\
&N; V << N\} \\
&without\{N2[upos = ADJ|ADV|INTJ|NOUN|PROPN| \\
&VERB|ADP|AUX|CCONJ|DET|NUM|PART|PRON| \\
&SCONJ|SYM]; N - [conj] - > N2\} \\
&without\{N2[upos = NOUN|PROPN]; N - [appos] - > N2\}
\end{aligned}
\tag{21.460}$$

(21.460) captures verbs with singular number before the nominal subjects or objects with singular number.

$$\begin{aligned}
&\%33 - VN - general - VsinNsin - Vaft \\
&pattern\{V[upos = VERB, Number = Sing]; N[upos = \\
&NOUN|PROPN, Number = Sing]; V - [comp : obj|subj] - > \\
&N; V >> N\} \\
&without\{N2[upos = ADJ|ADV|INTJ|NOUN|PROPN| \\
&VERB|ADP|AUX|CCONJ|DET|NUM|PART|PRON| \\
&SCONJ|SYM]; N - [conj] - > N2\} \\
&without\{N2[upos = NOUN|PROPN]; N - [appos] - > N2\}
\end{aligned}
\tag{21.461}$$

(21.461) captures verbs with singular number after the nominal subjects or objects with singular number.

$$\begin{aligned}
&\%33 - VN - general - VplNsin - Vbef \\
&pattern\{V[upos = VERB, Number = Plur]; N[upos = \\
&NOUN|PROPN, Number = Sing]; V - [comp : obj|subj] - > \\
&N; V << N\} \\
&without\{N2[upos = ADJ|ADV|INTJ|NOUN|PROPN| \\
&VERB|ADP|AUX|CCONJ|DET|NUM|PART|PRON| \\
&SCONJ|SYM]; N - [conj] - > N2\} \\
&without\{N2[upos = NOUN|PROPN]; N - [appos] - > N2\}
\end{aligned}
\tag{21.462}$$

(21.462) captures verbs with plural number before the nominal subjects or objects with singular number.

$$\begin{aligned}
& \%33 - VN - general - VplNsin - Vaft \\
& pattern\{V[upos = VERB, Number = Plur]; N[upos = \\
& NOUN|PROPN, Number = Sing]; V - [comp : obj|subj] - > \\
& N; V >> N\} \\
& without\{N2[upos = ADJ|ADV|INTJ|NOUN|PROPN| \\
& VERB|ADP|AUX|CCONJ|DET|NUM|PART|PRON| \\
& SCONJ|SYM]; N - [conj] - > N2\} \\
& without\{N2[upos = NOUN|PROPN]; N - [appos] - > N2\}
\end{aligned} \tag{21.463}$$

(21.463) captures verbs with plural number after the nominal subjects or objects with singular number.

$$\begin{aligned}
& \%33 - VN - general - VsinNpl - Vbef \\
& pattern\{V[upos = VERB, Number = Sing]; N[upos = \\
& NOUN|PROPN, Number = Plur]; V - [comp : obj|subj] - > \\
& N; V << N\} \\
& without\{N2[upos = ADJ|ADV|INTJ|NOUN|PROPN| \\
& VERB|ADP|AUX|CCONJ|DET|NUM|PART|PRON| \\
& SCONJ|SYM]; N - [conj] - > N2\} \\
& without\{N2[upos = NOUN|PROPN]; N - [appos] - > N2\}
\end{aligned} \tag{21.464}$$

(21.464) captures verbs with singular number before the nominal subjects or objects with plural number.

$$\begin{aligned}
& \%33 - VN - general - VsinNpl - Vaft \\
& pattern\{V[upos = VERB, Number = Sing]; N[upos = \\
& NOUN|PROPN, Number = Plur]; V - [comp : obj|subj] - > \\
& N; V >> N\} \\
& without\{N2[upos = ADJ|ADV|INTJ|NOUN|PROPN| \\
& VERB|ADP|AUX|CCONJ|DET|NUM|PART|PRON| \\
& SCONJ|SYM]; N - [conj] - > N2\} \\
& without\{N2[upos = NOUN|PROPN]; N - [appos] - > N2\}
\end{aligned} \tag{21.465}$$

(21.465) captures verbs with singular number after the nominal subjects or objects with plural number.

Subsequently, as mentioned above, we want to test the behavior of subjects and objects separately. Again, splitting the already proposed patterns and focusing on one dependency instead of both simultaneously will be sufficient. We take as an example the pattern (21.448), which we convert into (21.466) for subjects and (21.467) for objects. We do the same with the rest of the patterns.

$$\begin{array}{l}
 \%33 - VN - subj - nofeatVN - Vbef \\
 pattern\{V[upos = VERB, !Number]; N[upos = \\
 NOUN|PROPN, !Number]; V - [subj] - > N; V << N\} \\
 without\{N2[upos = ADJ|ADV|INTJ|NOUN|PROPN| \\
 VERB|ADP|AUX|CCONJ|DET|NUM|PART|PRON| \\
 SCONJ|SYM]; N - [conj] - > N2\} \\
 without\{N2[upos = NOUN|PROPN]; N - [appos] - > N2\}
 \end{array}
 \tag{21.466}$$

(21.466) captures verbs without number before nominal subjects without number.

$$\begin{array}{l}
 \%33 - VN - obj - nofeatVN - Vbef \\
 pattern\{V[upos = VERB, !Number]; N[upos = \\
 NOUN|PROPN, !Number]; V - [comp : obj] - > N; V << N\} \\
 without\{N2[upos = ADJ|ADV|INTJ|NOUN|PROPN| \\
 VERB|ADP|AUX|CCONJ|DET|NUM|PART|PRON| \\
 SCONJ|SYM]; N - [conj] - > N2\} \\
 without\{N2[upos = NOUN|PROPN]; N - [appos] - > N2\}
 \end{array}
 \tag{21.467}$$

(21.467) captures verbs without number before nominal objects without number.

21.33.2 Limitations Based on Grew-Match Data

As suspected when formalizing this Universal, it is impossible to obtain reliable results from this formulation. If one does not have a high knowledge of the functioning of each of the languages, it is impossible to know whether the number agreement between the noun and verb is suspended and that the language opts for using a rule based on the order.

21.34 Universal 34

UNIVERSAL 34: *No language has a trial number unless it has a dual. No language has a dual unless it has a plural.*

21.34.1 Extracting the Information in the Treebank Hub Page

This Universal can be checked via treebank hub page, since we have just to confirm the presence of an individual feature tag (if another one is also present). For this, we go to the treebank hub page of every corpus and select “Features>Number”. At the beginning of the chapter, it is said how many number values we find active in this corpus. In the case of Spanish (AnCora), there are two possible values: Plur (plural) and Sing (singular). Then, Spanish is not a language that must be checked to confirm this Universal since it does not have trials or duals.

21.34.2 Formalization Proposed for Grew-Match

Checking this Universal by Grew-Match is very simple. We simply create patterns in which the element we seek has “Number=” and then change the desired number value. In addition to checking the three explicit values in this hierarchy, we decided to check all possible features of number: trial, dual, plural, singular, paucal, greater paucal, greater plural, inverse number, count plural, plural tantum, and collective/mass singulare tantum. In this way, we may be able to refine further or extend the hierarchy proposed by Greenberg. These features have been previously explained in the abbreviations guide. We can present the formalizations (21.468) - (21.478).

$\begin{aligned} \%34 - TRI \\ pattern\{TRI[Number = Tri]\} \end{aligned}$	(21.468)

(21.468) captures trial numbers.

$\begin{aligned} \%34 - DUAL \\ pattern\{DUAL[Number = Dual]\} \end{aligned}$	(21.469)

(21.469) captures dual numbers.

$\begin{array}{l} \%34 - PLUR \\ pattern\{PLUR[Number = Plur]\} \end{array}$	(21.470)

(21.470) captures plural numbers.

$\begin{array}{l} \%34 - SING \\ pattern\{SING[Number = Sing]\} \end{array}$	(21.471)

(21.471) captures singular numbers.

$\begin{array}{l} \%34 - PAUC \\ pattern\{PAUC[Number = Pauc]\} \end{array}$	(21.472)

(21.472) captures paucal numbers.

$\begin{array}{l} \%34 - Grpa \\ pattern\{Grpa[Number = Grpa]\} \end{array}$	(21.473)

(21.473) captures greater paucal numbers.

$\begin{array}{l} \%34 - Grpl \\ pattern\{Grpl[Number = Grpl]\} \end{array}$	(21.474)

(21.474) captures greater plural numbers.

$\begin{array}{l} \%34 - Inv \\ pattern\{Inv[Number = Inv]\} \end{array}$	(21.475)

(21.475) captures inverse numbers.

$\begin{array}{l} \%34 - Count \\ pattern\{Count[Number = Count]\} \end{array}$	(21.476)

(21.476) captures count plural numbers.

$\begin{array}{l} \%34 - Ptan \\ pattern\{Ptan[Number = Ptan]\} \end{array}$	(21.477)

(21.477) captures plurale tantum numbers.

$$\begin{array}{l} \%34 - Coll \\ pattern\{Coll[Number = Coll]\} \end{array}$$

(21.478)

(21.478) captures collective numbers.

21.34.3 Trichotomous Expression Based on Grew-Match Data

First of all, we should mention that we have not found occurrences of words with trial numbers in the set available in UD, which causes the validity of the Universal can only be based on the second part (referring to dual-plural). However, it shows results of 100% validity. We also wanted to check whether the rest of the possible labels of number always appear in languages that have plural. Indeed, Coll (collective), Ptan (plurale tantum), and Coll (singulare tantum) appear in 100% of cases where there is also a plural. It is, therefore, possible to carry out an extension of the Universal in which all these categories are collected as numbers in the background, behind the plural.

We can represent the data in Fuzzy Logic as in figure 21.70.

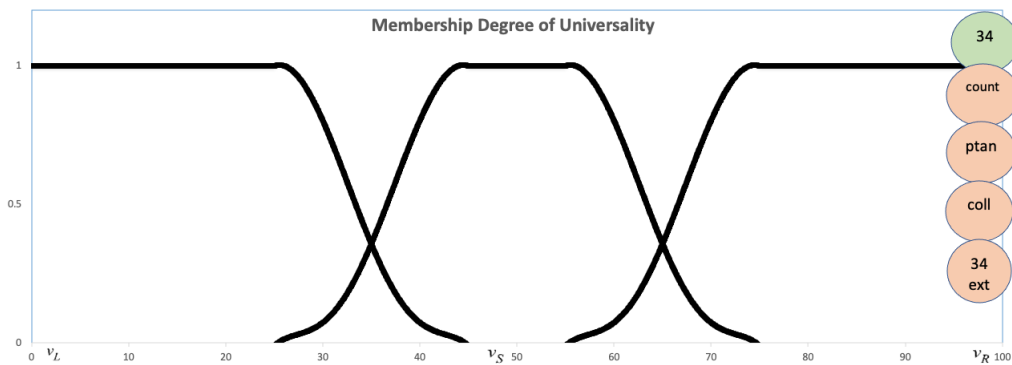


Figure 21.70: Trichotomous Expression for Universal 34

In this case, both Greenberg’s (dual) formulation, the others made by us, and the possible joint proposal of extending the Universal correspond to a prototypical high Universal. Such a claim, therefore, seems to be very illustrative of the functioning of languages in general. However, as with Universals with a closer approximation to deep Universals such as this one, its use in classifying languages is scarce (since its purpose is definitional). We can summarise our findings as in table 21.35.

Summary of characteristics of the Universal 34 analysis: 1. No attested trial numbers in our UD version. 2. The original Universal in dual-plural opposition is 100% accomplished. 3. The original Universal that we could check is a High-Universal. 4. We can extend the Universal by changing “dual” by “count”, “ptan”, or “coll” and it is also a High-Universal (100% accomplishment). Also, if we take all these categories at the same time.

Table 21.35: Summary of Characteristics of the Universal 34 Analysis

21.35 Universal 35

UNIVERSAL 35: *There is no language in which the plural does not have some non-zero allomorphs, whereas there are languages in which the singular is expressed only by zero. The dual and the trial are rarely expressed only by zero.*

21.35.1 Current Limitation: Data no Available

In this Universal, Greenberg is relating a value that we can know (number in languages) to another that we cannot know (allomorphs). The relationship and distribution between the different features of numbers and allomorphs cannot be known because we have no information about the latter in Universal Dependencies.

21.36 Universal 36

UNIVERSAL 36: *If a language has the gender category, it always has the number category.*

21.36.1 Extracting the Information in the Treebank Hub Page

This Universal can be checked via treebank hub page, since we have just to confirm the presence of an individual feature tag (if another one is also present). For this, we go to the treebank hub

page of every corpus and check if in “Features” there are the “gender” and “number” labels. It is also possible to mark any feature languages exhibited to check other correlations. For example, Spanish (AnCora) has gender and number, and Yoruba (YTB) only number.

21.36.2 Formalization Proposed for Grew-Match

Verification of this Universal via Grew-Match is, again, straightforward. We must create a pattern with a single element without any dependencies. In the case of the Universal, this element must be anything that exhibits the desired feature, a pattern that captures occurrences with the value of gender in any part of speech and another with a number. However, as we always want, we propose to extend this research to other features. In this case, we propose analyzing features applicable to the same type of words: animacy, case, definiteness, and degree. We will check whether these features also show some invariance similar to that proposed by Greenberg for a possible refinement of the Universal. In short, the patterns we have created are (21.479) - (21.484):

$\begin{array}{l} \%36 - \textit{Gender} \\ \textit{pattern}\{\textit{Gender}[\textit{Gender}]\} \end{array}$	(21.479)

(21.479) captures the presence of any gender.

$\begin{array}{l} \%36 - \textit{Number} \\ \textit{pattern}\{\textit{Number}[\textit{Number}]\} \end{array}$	(21.480)

(21.480) captures the presence of any number.

$\begin{array}{l} \%36 - \textit{Animacy} \\ \textit{pattern}\{\textit{Animacy}[\textit{Animacy}]\} \end{array}$	(21.481)

(21.481) captures the presence of any animacy.

$\begin{array}{l} \%36 - \textit{Case} \\ \textit{pattern}\{\textit{Case}[\textit{Case}]\} \end{array}$	(21.482)

(21.482) captures the presence of any case.

$\begin{array}{l} \%36 - \textit{Definite} \\ \textit{pattern}\{\textit{Definite}[\textit{Definite}]\} \end{array}$	(21.483)

(21.483) captures the presence of any definiteness.

$$\boxed{\begin{array}{l} \%36 - Degree \\ pattern\{Degree[Degree]\} \end{array}} \quad (21.484)$$

(21.484) captures the presence of any degree.

21.36.3 Trichotomous Expression Based on Grew-Match Data

In this Universal, we have been able to review Greenberg’s original formulation, extensions or modifications of the Universal, and a hierarchy that we will see later, although we will be checking it in the future.

About the Universal, as it has been conceived, we found compliance in 100% of the cases, which invites us to confirm the primacy of number over gender.

We first carried out individual checks following the same implicational structure to check the importance of number over other types of features. With animacy, we also found a 100% compliance. With case we found 99% compliance. With definiteness and degree, we found 98% compliance.

A formulation of Universal in which languages have all these features simultaneously also offers 96% confirmation. This invites us to formulate such an extension of the Universal, which represents many more features at once (and allows individual readings as above).

The location of this data on the Fuzzy Logic-based scale can be seen in figure 21.71.

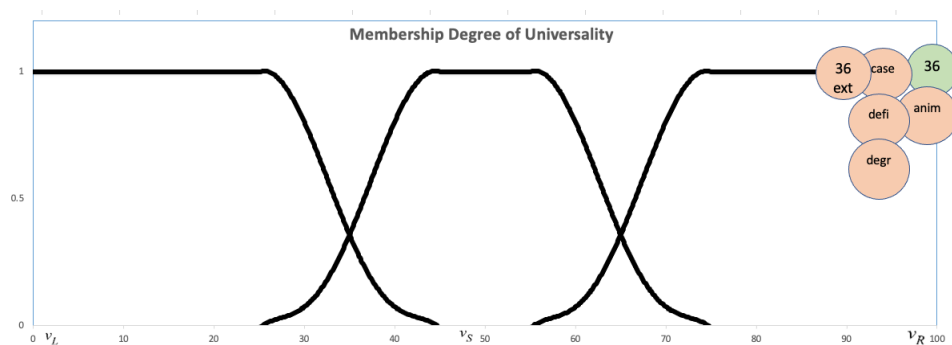


Figure 21.71: Trichotomous Expression for Universal 36

Greenberg’s original Universal is a prototypical High-Universal. So is the Universal we propose as an extension as well as the different relations between different categories and number. Therefore,

any formulations presented here are valid and of linguistic interest. We can summarise our findings as in table 21.36.

Summary of characteristics of the Universal 36 analysis:

1. There is a primacy of number over gender. with Greenberg's original Universal is 100 %. is also 100 % compliance if we extend the Universal to the animacy feature. is 99% compliance if we extend it to cases and 98% compliance in the case of definiteness. that have animacy, case, definiteness, and gender, in 96% of cases, will have a number.
2. Greenberg's original Universal and the checks we performed are High-Universal.

Table 21.36: Summary of Characteristics of the Universal 36 Analysis

21.37 Universal 37

UNIVERSAL 37: *A language never has more gender categories in non-singular numbers than in the singular.*

21.37.1 Formalization Proposed for Grew-Match

To check this Universal, we must check the occurrences in the languages of different combinations of gender and number features. Therefore, we have to check the occurrences of masculine with singulars, plurals, and the rest of the less extended numbers. Then, do the same with feminine, neuter, and common. That is, we must formalize as in:

$$\begin{array}{l} \%37 - FEM - SG \\ pattern\{FEMSG[Gender = Fem, Number = Sing]\} \end{array} \quad (21.485)$$

(21.485) captures languages with occurrences with feminine gender and singular number.

$$\begin{array}{l} \%37 - FEM - PL \\ pattern\{FEMPL[Gender = Fem, Number = Plur]\} \end{array} \quad (21.486)$$

(21.486) captures languages with occurrences with feminine gender and plural number.

$$\begin{array}{l} \%37 - MASC - SG \\ pattern\{MASC SG[Gender = Masc, Number = Sing]\} \end{array} \quad (21.487)$$

(21.487) captures languages with occurrences with masculine gender and singular number.

$$\begin{array}{l} \%37 - MASC - PL \\ pattern\{MASC PL[Gender = Masc, Number = Plur]\} \end{array} \quad (21.488)$$

(21.488) captures languages with occurrences with masculine gender and plural number.

$$\begin{array}{l} \%37 - Neut - SG \\ pattern\{NEUT SG[Gender = Neut, Number = Sing]\} \end{array} \quad (21.489)$$

(21.489) captures languages with occurrences with neutral gender and singular number.

$$\begin{array}{l} \%37 - Neut - PL \\ pattern\{NEUT PL[Gender = Neut, Number = Plur]\} \end{array} \quad (21.490)$$

(21.490) captures languages with occurrences with neutral gender and plural number.

$$\begin{array}{l} \%37 - Com - SG \\ pattern\{COM SG[Gender = Com, Number = Sing]\} \end{array} \quad (21.491)$$

(21.491) captures languages with occurrences with common gender and singular number.

$$\begin{array}{l} \%37 - Com - PL \\ pattern\{COM PL[Gender = Com, Number = Plur]\} \end{array} \quad (21.492)$$

(21.492) captures languages with occurrences with common gender and plural number.

$$\begin{array}{l} \%37 - MASC - REST \\ pattern\{MASC REST[Gender = Masc, Number = \\ Dual|Tri|Pauc|Grpa|Grpl|Inv|Count|Ptan|Coll]\} \end{array} \quad (21.493)$$

(21.493) captures languages with occurrences with masculine gender and dual, trial, paucal, greater paucal, greater plural, inverse, count plural, plurale tantum, or collective number.

$$\begin{array}{l}
\%37 - FEM - REST \\
pattern\{FEMREST[Gender = Fem, Number = \\
Dual|Tri|Pauc|Grpa|Grpl|Inv|Count|Ptan|Coll]\}
\end{array} \quad (21.494)$$

(21.494) captures languages with occurrences with masculine gender and dual, trial, paucal, greater paucal, greater plural, inverse, count plural, plurale tantum, or collective number.

$$\begin{array}{l}
\%37 - Com - REST \\
pattern\{COMREST[Gender = Com, Number = \\
Dual|Tri|Pauc|Grpa|Grpl|Inv|Count|Ptan|Coll]\}
\end{array} \quad (21.495)$$

(21.495) captures languages with occurrences with common gender and dual, trial, paucal, greater paucal, greater plural, inverse, count plural, plurale tantum, or collective number.

$$\begin{array}{l}
\%37 - Neut - REST \\
pattern\{NEUTREST[Gender = Neut, Number = \\
Dual|Tri|Pauc|Grpa|Grpl|Inv|Count|Ptan|Coll]\}
\end{array} \quad (21.496)$$

(21.496) captures languages with occurrences with neutral gender and dual, trial, paucal, greater paucal, greater plural, inverse, count plural, plurale tantum, or collective number.

As can be seen, the patterns simply collect all possible combinations of genders with different numbers. For example, we have named the element *FEMSG* as any part of speech with feminine gender and singular number. In the case of less frequent numerals, we have grouped their search since it is more than enough to check the Universal, and the specific occurrences we may obtain would not be significant. For the Universal to be true, we must look at occurrences of the different genders in singular and find that there are more genders (or the same) in singular than genders with occurrences in plural or the grouping of other genders.

21.37.2 Trichotomous Expression Based on Grew-Match Data

The fulfillment of this Universal is manifested in 100% of the possible languages. Its representation on a trichotomous scale can be seen in the figure 21.72.

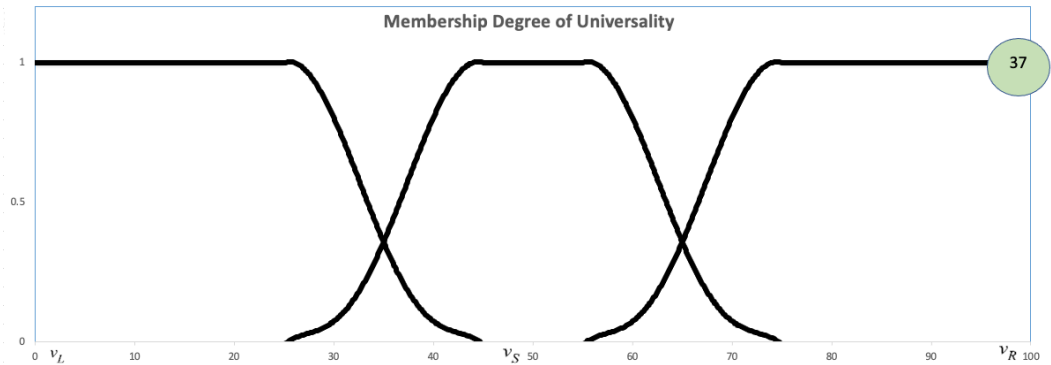


Figure 21.72: Trichotomous Expression for Universal 37

This quantitatively strong result corresponds to a prototypical High-Universal. Its location on the scale used in this doctoral thesis allows us to determine that, even if some languages were to appear as exceptions to this Universal, the importance of this Universal would not change at all. In other words, its predictive value and typological interest are high. We can summarise our findings as in table 21.37.

Summary of characteristics of the Universal 37 analysis:
1. Greenberg’s original Universal is fulfilled in 100% of the cases.
2. Greenberg’s original Universal is a High-Universal.

Table 21.37: Summary of Characteristics of the Universal 37 Analysis

21.38 Universal 38

UNIVERSAL 38: *Where there is a case system, the only case which ever has only zero allomorphs is the one which includes that of the subject of the intransitive verb among its meanings.*

21.38.1 Current Limitation: Data no Available

In this Universal case, Greenberg again mixes an element that we know with the same element that we cannot know. That is, he establishes links between specific cases (in case they exist in the languages analyzed) and the presence or not of allomorphs.

21.39 Universal 39

UNIVERSAL 39: *Where morphemes of both number and case are present and both follow, or both precede the noun base, the expression of number almost always comes between the noun base and the expression of case.*

21.39.1 Current Limitation: Data no Available

In the case of Universal 39, Greenberg tries to systematize the order of number and case morphemes when they appear simultaneously and in the same direction as the root. We do not know completely the location of these morphemes in the world’s languages in UD; we only know whether a word carries this information. Again, in this case, the use of interglossae would be very beneficial and effective since both morphemes would be isolated, and their order could be computed.

21.40 Universal 40

UNIVERSAL 40: *When the adjective follows the noun, the adjective expresses all the inflectional categories of the noun. In such cases the noun may lack over expression of one or all of these categories.*

21.40.1 Formalization Proposed for Grew-Match

In order to check this Universal, we must consider different linguistic aspects that must be combined to obtain the occurrences. First, we observe a word order distinction between the depending adjective and the noun head. Therefore, formalizations with the possibility of this adjective being in the first or second position must be considered. Subsequently, Greenberg combines this position with the expression of the inflectional categories of the noun and the adjective without detailing which ones (whichever they are). These categories can be, according to UD: gender, animacy, noun class, number, case, definite, and degree. In this case, we must combine word orders and the absence or presence of these features in the two elements for which a dependency has been created, with a noun as head and an adjective as dependent. For example, about gender, both can have (possible according to Greenberg), both cannot have (possible according to Greenberg)

but not checked in this Universal), the adjective can have, and the noun cannot have (possible according to Greenberg), the adjective cannot have, and the noun can have (impossible according to Greenberg). This restriction applies only to adjectives after nouns, according to Greenberg; we will also check the behavior of adjectives before nouns.

It is also interesting to see which languages opt for the efficiency option (the noun does not need to express the features if the adjective already does) and which languages prefer harmony (both words express the same thing). In addition, it will also be interesting to review, in a detailed and precise way, feature by feature to see if it applies to all cases or if some inflectional features are much more prone to it.

In summary, we propose the patterns (21.497) - (21.544).

$$\begin{array}{l} \%40 - \textit{Gender} - \textit{BOTH}, \textit{NADJ} \\ \textit{pattern}\{\textit{ADJ}[\textit{upos} = \textit{ADJ}, \textit{Gender}]; \textit{NOUN}[\textit{upos} = \\ \textit{NOUN}, \textit{Gender}]; \textit{NOUN} - > \textit{ADJ}; \textit{NOUN} << \textit{ADJ}\} \end{array} \quad (21.497)$$

(21.497) captures adjectives with any gender after nouns with any gender value.

$$\begin{array}{l} \% - 40 - \textit{Gender} - \textit{BOTH}, \textit{ADJN} \\ \textit{pattern}\{\textit{ADJ}[\textit{upos} = \textit{ADJ}, \textit{Gender}]; \textit{NOUN}[\textit{upos} = \\ \textit{NOUN}, \textit{Gender}]; \textit{NOUN} - > \textit{ADJ}; \textit{NOUN} >> \textit{ADJ}\} \end{array} \quad (21.498)$$

(21.498) captures adjectives with any gender before nouns with any gender value.

$$\begin{array}{l} \%40 - \textit{Gender} - \textit{ADJ}, \textit{NADJ} \\ \textit{pattern}\{\textit{ADJ}[\textit{upos} = \textit{ADJ}, \textit{Gender}]; \textit{NOUN}[\textit{upos} = \\ \textit{NOUN}, !\textit{Gender}]; \textit{NOUN} - > \textit{ADJ}; \textit{NOUN} << \textit{ADJ}\} \end{array} \quad (21.499)$$

(21.499) captures adjectives with any gender after nouns without any gender value.

$$\begin{array}{l} \%40 - \textit{Gender} - \textit{ADJ}, \textit{ADJN} \\ \textit{pattern}\{\textit{ADJ}[\textit{upos} = \textit{ADJ}, \textit{Gender}]; \textit{NOUN}[\textit{upos} = \\ \textit{NOUN}, !\textit{Gender}]; \textit{NOUN} - > \textit{ADJ}; \textit{NOUN} >> \textit{ADJ}\} \end{array} \quad (21.500)$$

(21.500) captures adjectives with any gender before nouns without any gender value.

$$\begin{array}{l}
\%40 - \textit{Gender} - \textit{NOUN}, \textit{NADJ} \\
\textit{pattern}\{\textit{ADJ}[\textit{upos}] = \textit{ADJ}, !\textit{Gender}; \textit{NOUN}[\textit{upos}] = \\
\textit{NOUN}, \textit{Gender}; \textit{NOUN} - > \textit{ADJ}; \textit{NOUN} << \textit{ADJ}\}
\end{array} \quad (21.501)$$

(21.501) captures adjectives without any gender after nouns with any gender value.

$$\begin{array}{l}
\%40 - \textit{Gender} - \textit{NOUN}, \textit{ADJN} \\
\textit{pattern}\{\textit{ADJ}[\textit{upos}] = \textit{ADJ}, !\textit{Gender}; \textit{NOUN}[\textit{upos}] = \\
\textit{NOUN}, \textit{Gender}; \textit{NOUN} - > \textit{ADJ}; \textit{NOUN} >> \textit{ADJ}\}
\end{array} \quad (21.502)$$

(21.502) captures adjectives without any gender before nouns with any gender value.

$$\begin{array}{l}
\%40 - \textit{Animacy} - \textit{BOTH}, \textit{NADJ} \\
\textit{pattern}\{\textit{ADJ}[\textit{upos}] = \textit{ADJ}, \textit{Animacy}; \textit{NOUN}[\textit{upos}] = \\
\textit{NOUN}, \textit{Animacy}; \textit{NOUN} - > \textit{ADJ}; \textit{NOUN} << \textit{ADJ}\}
\end{array} \quad (21.503)$$

(21.503) captures adjectives with any animacy after nouns with any animacy value.

$$\begin{array}{l}
\%40 - \textit{Animacy} - \textit{BOTH}, \textit{ADJN} \\
\textit{pattern}\{\textit{ADJ}[\textit{upos}] = \textit{ADJ}, \textit{Animacy}; \textit{NOUN}[\textit{upos}] = \\
\textit{NOUN}, \textit{Animacy}; \textit{NOUN} - > \textit{ADJ}; \textit{NOUN} >> \textit{ADJ}\}
\end{array} \quad (21.504)$$

(21.504) captures adjectives with any animacy before nouns with any animacy value.

$$\begin{array}{l}
\%40 - \textit{Animacy} - \textit{ADJ}, \textit{NADJ} \\
\textit{pattern}\{\textit{ADJ}[\textit{upos}] = \textit{ADJ}, \textit{Animacy}; \textit{NOUN}[\textit{upos}] = \\
\textit{NOUN}, !\textit{Animacy}; \textit{NOUN} - > \textit{ADJ}; \textit{NOUN} << \textit{ADJ}\}
\end{array} \quad (21.505)$$

(21.505) captures adjectives with any animacy after nouns without any animacy value.

$$\begin{array}{l}
\%40 - \textit{Animacy} - \textit{ADJ}, \textit{ADJN} \\
\textit{pattern}\{\textit{ADJ}[\textit{upos}] = \textit{ADJ}, \textit{Animacy}; \textit{NOUN}[\textit{upos}] = \\
\textit{NOUN}, !\textit{Animacy}; \textit{NOUN} - > \textit{ADJ}; \textit{NOUN} >> \textit{ADJ}\}
\end{array} \quad (21.506)$$

(21.506) captures adjectives with any animacy before nouns without any animacy value.

$$\begin{array}{l}
 \%40 - \textit{Animacy} - \textit{NOUN}, \textit{NADJ} \\
 \textit{pattern}\{\textit{ADJ}[\textit{upos} = \textit{ADJ}, !\textit{Animacy}]; \textit{NOUN}[\textit{upos} = \\
 \textit{NOUN}, \textit{Animacy}]; \textit{NOUN} - > \textit{ADJ}; \textit{NOUN} << \textit{ADJ}\}
 \end{array} \quad (21.507)$$

(21.507) captures adjectives without any animacy after nouns with any animacy value.

$$\begin{array}{l}
 \%40 - \textit{Animacy} - \textit{NOUN}, \textit{ADJN} \\
 \textit{pattern}\{\textit{ADJ}[\textit{upos} = \textit{ADJ}, !\textit{Animacy}]; \textit{NOUN}[\textit{upos} = \\
 \textit{NOUN}, \textit{Animacy}]; \textit{NOUN} - > \textit{ADJ}; \textit{NOUN} >> \textit{ADJ}\}
 \end{array} \quad (21.508)$$

(21.508) captures adjectives without any animacy before nouns with any animacy value.

$$\begin{array}{l}
 \%40 - \textit{NounClass} - \textit{BOTH}, \textit{NADJ} \\
 \textit{pattern}\{\textit{ADJ}[\textit{upos} = \textit{ADJ}, \textit{NounClass}]; \textit{NOUN}[\textit{upos} = \\
 \textit{NOUN}, \textit{NounClass}]; \textit{NOUN} - > \textit{ADJ}; \textit{NOUN} << \textit{ADJ}\}
 \end{array} \quad (21.509)$$

(21.509) captures adjectives with any noun class after nouns with any noun class value.

$$\begin{array}{l}
 \%40 - \textit{NounClass} - \textit{BOTH}, \textit{ADJN} \\
 \textit{pattern}\{\textit{ADJ}[\textit{upos} = \textit{ADJ}, \textit{NounClass}]; \textit{NOUN}[\textit{upos} = \\
 \textit{NOUN}, \textit{NounClass}]; \textit{NOUN} - > \textit{ADJ}; \textit{NOUN} >> \textit{ADJ}\}
 \end{array} \quad (21.510)$$

(21.510) captures adjectives with any noun class before nouns with any noun class value.

$$\begin{array}{l}
 \%40 - \textit{NounClass} - \textit{ADJ}, \textit{NADJ} \\
 \textit{pattern}\{\textit{ADJ}[\textit{upos} = \textit{ADJ}, \textit{NounClass}]; \textit{NOUN}[\textit{upos} = \\
 \textit{NOUN}, !\textit{NounClass}]; \textit{NOUN} - > \textit{ADJ}; \textit{NOUN} << \textit{ADJ}\}
 \end{array} \quad (21.511)$$

(21.511) captures adjectives with any noun class value after nouns without any noun class value.

$$\begin{array}{l}
 \%40 - \textit{NounClass} - \textit{ADJ}, \textit{ADJN} \\
 \textit{pattern}\{\textit{ADJ}[\textit{upos} = \textit{ADJ}, \textit{NounClass}]; \textit{NOUN}[\textit{upos} = \\
 \textit{NOUN}, !\textit{NounClass}]; \textit{NOUN} - > \textit{ADJ}; \textit{NOUN} >> \textit{ADJ}\}
 \end{array} \quad (21.512)$$

(21.512) captures adjectives with any noun class value before nouns without any noun class value.

$$\begin{array}{l}
\%40 - NounClass - NOUN, NADJ \\
pattern\{ADJ[upos = ADJ, !NounClass]; NOUN[upos = \\
NOUN, NounClass]; NOUN - > ADJ; NOUN << ADJ\}
\end{array} \quad (21.513)$$

(21.513) captures adjectives without any noun class value after nouns with any noun class value.

$$\begin{array}{l}
\%40 - NounClass - NOUN, ADJN \\
pattern\{ADJ[upos = ADJ, !NounClass]; NOUN[upos = \\
NOUN, NounClass]; NOUN - > ADJ; NOUN >> ADJ\}
\end{array} \quad (21.514)$$

(21.514) captures adjectives without any noun class value before nouns without any noun class value.

$$\begin{array}{l}
\%40 - Number - BOTH, NADJ \\
pattern\{ADJ[upos = ADJ, Number]; NOUN[upos = \\
NOUN, Number]; NOUN - > ADJ; NOUN << ADJ\}
\end{array} \quad (21.515)$$

(21.515) captures adjectives with any number value after nouns with any number value.

$$\begin{array}{l}
\%40 - Number - BOTH, ADJN \\
pattern\{ADJ[upos = ADJ, Number]; NOUN[upos = \\
NOUN, Number]; NOUN - > ADJ; NOUN >> ADJ\}
\end{array} \quad (21.516)$$

(21.516) captures adjectives with any number value before nouns with any number value.

$$\begin{array}{l}
\%40 - Number - ADJ, NADJ \\
pattern\{ADJ[upos = ADJ, Number]; NOUN[upos = \\
NOUN, !Number]; NOUN - > ADJ; NOUN << ADJ\}
\end{array} \quad (21.517)$$

(21.517) captures adjectives with any number value after nouns without any number value.

$$\begin{array}{l}
\%40 - Number - ADJ, ADJN \\
pattern\{ADJ[upos = ADJ, Number]; NOUN[upos = \\
NOUN, !Number]; NOUN - > ADJ; NOUN >> ADJ\}
\end{array} \quad (21.518)$$

(21.518) captures adjectives with any number value before nouns without any number value.

$$\begin{array}{l} \%40 - \textit{Number} - \textit{NOUN}, \textit{NADJ} \\ \textit{pattern}\{\textit{ADJ}[\textit{upos}] = \textit{ADJ}, !\textit{Number}\}; \textit{NOUN}[\textit{upos}] = \\ \textit{NOUN}, \textit{Number}\}; \textit{NOUN} - > \textit{ADJ}; \textit{NOUN} << \textit{ADJ}\} \end{array} \quad (21.519)$$

(21.519) captures adjectives without any number value after nouns with any number value.

$$\begin{array}{l} \%40 - \textit{Number} - \textit{NOUN}, \textit{ADJN} \\ \textit{pattern}\{\textit{ADJ}[\textit{upos}] = \textit{ADJ}, !\textit{Number}\}; \textit{NOUN}[\textit{upos}] = \\ \textit{NOUN}, \textit{Number}\}; \textit{NOUN} - > \textit{ADJ}; \textit{NOUN} >> \textit{ADJ}\} \end{array} \quad (21.520)$$

(21.520) captures adjectives without any number value before nouns with any number value.

$$\begin{array}{l} \%40 - \textit{Case} - \textit{BOTH}, \textit{NADJ} \\ \textit{pattern}\{\textit{ADJ}[\textit{upos}] = \textit{ADJ}, \textit{Case}\}; \textit{NOUN}[\textit{upos}] = \\ \textit{NOUN}, \textit{Case}\}; \textit{NOUN} - > \textit{ADJ}; \textit{NOUN} << \textit{ADJ}\} \end{array} \quad (21.521)$$

(21.521) captures adjectives with any case value after nouns with any case value.

$$\begin{array}{l} \%40 - \textit{Case} - \textit{BOTH}, \textit{ADJN} \\ \textit{pattern}\{\textit{ADJ}[\textit{upos}] = \textit{ADJ}, \textit{Case}\}; \textit{NOUN}[\textit{upos}] = \\ \textit{NOUN}, \textit{Case}\}; \textit{NOUN} - > \textit{ADJ}; \textit{NOUN} >> \textit{ADJ}\} \end{array} \quad (21.522)$$

(21.522) captures adjectives with any case value before nouns with any case value.

$$\begin{array}{l} \%40 - \textit{Case} - \textit{ADJ}, \textit{NADJ} \\ \textit{pattern}\{\textit{ADJ}[\textit{upos}] = \textit{ADJ}, \textit{Case}\}; \textit{NOUN}[\textit{upos}] = \\ \textit{NOUN}, !\textit{Case}\}; \textit{NOUN} - > \textit{ADJ}; \textit{NOUN} << \textit{ADJ}\} \end{array} \quad (21.523)$$

(21.523) captures adjectives with any case value after nouns without any case value.

$$\begin{array}{l} \%40 - \textit{Case} - \textit{ADJ}, \textit{ADJN} \\ \textit{pattern}\{\textit{ADJ}[\textit{upos}] = \textit{ADJ}, \textit{Case}\}; \textit{NOUN}[\textit{upos}] = \\ \textit{NOUN}, !\textit{Case}\}; \textit{NOUN} - > \textit{ADJ}; \textit{NOUN} >> \textit{ADJ}\} \end{array} \quad (21.524)$$

(21.524) captures adjectives with any case value before nouns without any case value.

$$\begin{array}{l}
\%40 - Case - NOUN, NADJ \\
pattern\{ADJ[upos = ADJ, !Case]; NOUN[upos = \\
NOUN, Case]; NOUN - > ADJ; NOUN << ADJ\}
\end{array} \quad (21.525)$$

(21.525) captures adjectives without any case value after nouns with any case value.

$$\begin{array}{l}
\%40 - Case - NOUN, ADJN \\
pattern\{ADJ[upos = ADJ, !Case]; NOUN[upos = \\
NOUN, Case]; NOUN - > ADJ; NOUN >> ADJ\}
\end{array} \quad (21.526)$$

(21.526) captures adjectives without any case value before nouns with any case value.

$$\begin{array}{l}
\%40 - Definite - BOTH, NADJ \\
pattern\{ADJ[upos = ADJ, Definite]; NOUN[upos = \\
NOUN, Definite]; NOUN - > ADJ; NOUN << ADJ\}
\end{array} \quad (21.527)$$

(21.527) captures adjectives with any definiteness value after nouns with any definiteness value.

$$\begin{array}{l}
\%40 - Definite - BOTH, ADJN \\
pattern\{ADJ[upos = ADJ, Definite]; NOUN[upos = \\
NOUN, Definite]; NOUN - > ADJ; NOUN >> ADJ\}
\end{array} \quad (21.528)$$

(21.528) captures adjectives with any definiteness value before nouns with any definiteness value.

$$\begin{array}{l}
\%40 - Definite - ADJ, NADJ \\
pattern\{ADJ[upos = ADJ, Definite]; NOUN[upos = \\
NOUN, !Definite]; NOUN - > ADJ; NOUN << ADJ\}
\end{array} \quad (21.529)$$

(21.529) captures adjectives with any definiteness value after nouns without any definiteness value.

$$\begin{array}{l}
\%40 - Definite - ADJ, ADJN \\
pattern\{ADJ[upos = ADJ, Definite]; NOUN[upos = \\
NOUN, !Definite]; NOUN - > ADJ; NOUN >> ADJ\}
\end{array} \quad (21.530)$$

(21.530) captures adjectives with any definiteness value before nouns without any definiteness value.

$$\begin{array}{l}
 \%40 - \textit{Definite} - \textit{NOUN}, \textit{NADJ} \\
 \textit{pattern}\{\textit{ADJ}[\textit{upos} = \textit{ADJ}, !\textit{Definite}]; \textit{NOUN}[\textit{upos} = \\
 \textit{NOUN}, \textit{Definite}]; \textit{NOUN} - > \textit{ADJ}; \textit{NOUN} << \textit{ADJ}\}
 \end{array} \quad (21.531)$$

(21.531) captures adjectives without any definiteness value after nouns with any definiteness value.

$$\begin{array}{l}
 \%40 - \textit{Definite} - \textit{NOUN}, \textit{ADJN} \\
 \textit{pattern}\{\textit{ADJ}[\textit{upos} = \textit{ADJ}, !\textit{Definite}]; \textit{NOUN}[\textit{upos} = \\
 \textit{NOUN}, \textit{Definite}]; \textit{NOUN} - > \textit{ADJ}; \textit{NOUN} >> \textit{ADJ}\}
 \end{array} \quad (21.532)$$

(21.532) captures adjectives without any definiteness value before nouns with any definiteness value.

$$\begin{array}{l}
 \%40 - \textit{Gender} - \textit{none}, \textit{NADJ} \\
 \textit{pattern}\{\textit{ADJ}[\textit{upos} = \textit{ADJ}, !\textit{Gender}]; \textit{NOUN}[\textit{upos} = \\
 \textit{NOUN}, !\textit{Gender}]; \textit{NOUN} - > \textit{ADJ}; \textit{NOUN} << \textit{ADJ}\}
 \end{array} \quad (21.533)$$

(21.533) captures adjectives without any gender value after nouns without any gender value.

$$\begin{array}{l}
 \%40 - \textit{Gender} - \textit{none}, \textit{ADJN} \\
 \textit{pattern}\{\textit{ADJ}[\textit{upos} = \textit{ADJ}, !\textit{Gender}]; \textit{NOUN}[\textit{upos} = \\
 \textit{NOUN}, !\textit{Gender}]; \textit{NOUN} - > \textit{ADJ}; \textit{NOUN} >> \textit{ADJ}\}
 \end{array} \quad (21.534)$$

(21.534) captures adjectives without any gender value before nouns without any gender value.

$$\begin{array}{l}
 \%40 - \textit{Animacy} - \textit{none}, \textit{NADJ} \\
 \textit{pattern}\{\textit{ADJ}[\textit{upos} = \textit{ADJ}, !\textit{Animacy}]; \textit{NOUN}[\textit{upos} = \\
 \textit{NOUN}, !\textit{Animacy}]; \textit{NOUN} - > \textit{ADJ}; \textit{NOUN} << \textit{ADJ}\}
 \end{array} \quad (21.535)$$

(21.535) captures adjectives without any animacy value after nouns without any animacy value.

$$\begin{array}{l}
 \%40 - \textit{Animacy} - \textit{none}, \textit{ADJN} \\
 \textit{pattern}\{\textit{ADJ}[\textit{upos} = \textit{ADJ}, !\textit{Animacy}]; \textit{NOUN}[\textit{upos} = \\
 \textit{NOUN}, !\textit{Animacy}]; \textit{NOUN} - > \textit{ADJ}; \textit{NOUN} >> \textit{ADJ}\}
 \end{array} \quad (21.536)$$

(21.536) captures adjectives without any animacy value before nouns without any animacy value.

$$\begin{array}{l}
\%40 - NounClass - none, NADJ \\
pattern\{ADJ[upos = ADJ, !NounClass]; NOUN[upos = \\
NOUN, !NounClass]; NOUN- > ADJ; NOUN << ADJ\}
\end{array} \quad (21.537)$$

(21.537) captures adjectives without any noun class value after nouns without any noun class value.

$$\begin{array}{l}
\%40 - NounClass - none, ADJN \\
pattern\{ADJ[upos = ADJ, !NounClass]; NOUN[upos = \\
NOUN, !NounClass]; NOUN- > ADJ; NOUN >> ADJ\}
\end{array} \quad (21.538)$$

(21.538) captures adjectives without any noun class value before nouns without any noun class value.

$$\begin{array}{l}
\%40 - Number - none, NADJ \\
pattern\{ADJ[upos = ADJ, !Number]; NOUN[upos = \\
NOUN, !Number]; NOUN- > ADJ; NOUN << ADJ\}
\end{array} \quad (21.539)$$

(21.539) captures adjectives without any number value after nouns without any number value.

$$\begin{array}{l}
\%40 - Number - none, ADJN \\
pattern\{ADJ[upos = ADJ, !Number]; NOUN[upos = \\
NOUN, !Number]; NOUN- > ADJ; NOUN >> ADJ\}
\end{array} \quad (21.540)$$

(21.540) captures adjectives without any number value before nouns without any number value.

$$\begin{array}{l}
\%40 - Case - none, NADJ \\
pattern\{ADJ[upos = ADJ, !Case]; NOUN[upos = \\
NOUN, !Case]; NOUN- > ADJ; NOUN << ADJ\}
\end{array} \quad (21.541)$$

(21.541) captures adjectives without any case value after nouns without any case value.

$$\begin{array}{l}
\%40 - Case - none, ADJN \\
pattern\{ADJ[upos = ADJ, !Case]; NOUN[upos = \\
NOUN, !Case]; NOUN- > ADJ; NOUN >> ADJ\}
\end{array} \quad (21.542)$$

(21.542) captures adjectives without any case value before nouns without any case value.

$$\begin{array}{l}
\%40 - \textit{Definite} - \textit{none}, \textit{NADJ} \\
\textit{pattern}\{\textit{ADJ}[\textit{upos} = \textit{ADJ}, !\textit{Definite}]; \textit{NOUN}[\textit{upos} = \\
\textit{NOUN}, !\textit{Definite}]; \textit{NOUN} - > \textit{ADJ}; \textit{NOUN} << \textit{ADJ}\}
\end{array} \quad (21.543)$$

(21.543) captures adjectives without any definiteness value after nouns without any definiteness value.

$$\begin{array}{l}
\%40 - \textit{Definite} - \textit{none}, \textit{ADJN} \\
\textit{pattern}\{\textit{ADJ}[\textit{upos} = \textit{ADJ}, !\textit{Definite}]; \textit{NOUN}[\textit{upos} = \\
\textit{NOUN}, !\textit{Definite}]; \textit{NOUN} - > \textit{ADJ}; \textit{NOUN} >> \textit{ADJ}\}
\end{array} \quad (21.544)$$

(21.544) captures adjectives without any definiteness value before nouns without any definiteness value.

In all cases, we activate two elements, one nominal and one adjectival, in which the latter depends on the former (as indicated by Greenberg). Subsequently, in each pattern, we add a different distribution of each feature between the two parts of speech: both have the feature, neither has it, the first has it and the second does not, the second has it, and the first does not. We formalize with the order constraint indicated by Greenberg “*NOUN* << *ADJ*”, and also with the opposite “*NOUN* >> *ADJ*”. We repeat this with each of the available features. Adding occurrences where none of the parts of speech contains a given feature is of no interest to the Universal itself. However, we have decided to add it to offer this information, which is interesting for the detailed knowledge of the languages and their linguistic types.

We expect very few occurrences of the NounClass feature. This is because we do not believe that it is a truly Universal tag, as it serves as a sort of catch-all for quite specific classificatory systems for specific languages (such as Wolof, for example).

21.40.2 Trichotomous Expression Based on Grew-Match Data

In this Universal, we also found some inconsistencies on the part of the labelers of the various treebanks. However, we have found that these are usually only a few specific examples that do not affect the entire individual type of each language. Moreover, they have been detected in a few languages and only in the gender category. For example, when we compare two very similar languages, such as Spanish and Portuguese, we obtain different results due to some gender annotation disparities. The adjective “radical”, which is written the same in Spanish, English, and

Portuguese, in Portuguese has the mark “fem” (because the noun in the context checked is fem). However, in Spanish, it does not have any gender marking (“!Gender”). This, in short, causes Spanish to be an exception to Universal 40 (the noun has gender, and the adjective does not). At the same time, Portuguese would comply with the Universal (both words contain gender features).

It is unusual to find in analyses such as the one we carry out revisions to errors of this type. We believe they are necessary to avoid biases and do not understand them as disregarding the enormous and valuable work of the UD treebanks scorers. We simply want to contribute to their further improvement. In short, the results we will show below will have to be contrasted once the proposed revisions are carried out to see if there is a significant deviation or, as we believe, the results would not vary significantly. Mainly because in most of the features tested we have not found this to be the case, even though we have always been able to overlook some errors.

1. The original Universal formulated by Greenberg generally has a 42% compliance rate.
2. We found 44% acceptance in the case of gender, 48% in the case of number, 56% in case, and 55% in definiteness.

Within the trichotomous model, these figures would be represented as in figure 21.73.

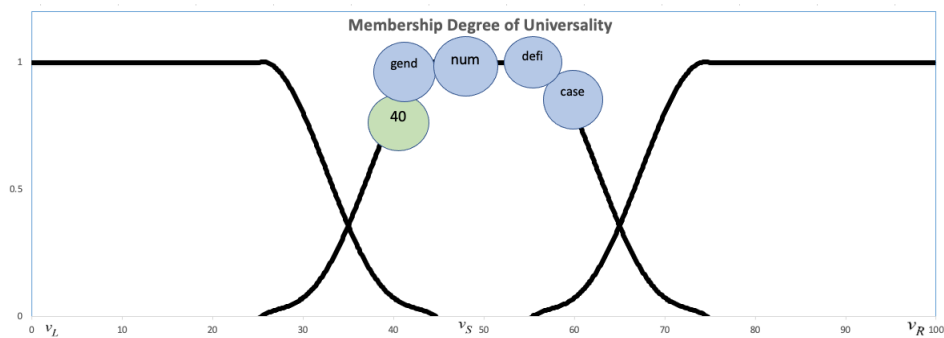


Figure 21.73: Trichotomous Expression for Universal 40

Both in the original and general formulation of the Universal and in the concrete analysis of the different detailed elements that contribute to it, we obtain Mid-Universals with quite high values of prototypicality. This means that this Universal should not be considered when working with languages because of its low predictive power since languages exhibit this Universal in more or less half of the cases. Another relevant aspect that we want to highlight about the model we use is that even if the concrete results of the analysis change in the future when better data become available, this shift will not be large enough (rarely) to change the set abruptly. This makes us

more confident in representing this Universal as a mere Mid-Universal. We can summarise our findings as in table 21.38.

Summary of characteristics of the Universal 40 analysis: 1. This is a good Universal for checking inconsistencies in annotation and similar problems detected. 2. The original Greenberg Universal is 42% compliant according to the data obtained. 3. The Universal comprises 44% compliance for gender, 48% compliance for number, 56% compliance for case, and 55% compliance for definiteness. 4. Greenberg’s original Universal and the individual checking of its parts are Mid-Universals. It does not seem that the direction of the adjective is decisive for the expression of the inflectional categories.

Table 21.38: Summary of Characteristics of the Universal 40 Analysis

21.41 Universal 41

UNIVERSAL 41: *If in a language, the verb follows both the nominal subject and nominal object as the dominant order, the language almost always has a case system.*

21.41.1 How to Check it Generating Tables in Grew-Match

We have decided to place this Universal in this section because it is a perfect occasion to demonstrate the use of the tables obtained in Grew-Match as a much more visual and direct alternative to formalization. However, to do so, we must first work with formulas directly in Grew-Match, already used in other Universals. First, we must check the basic dominant word order of the languages. To do this, we can reuse again the patterns of the Universal 1 (21.1): (21.1) - (21.6). Subsequently, as mentioned above, we need to determine dominance by ordering the frequency of occurrences of the different orders and assign the label of the most frequent order if it is twice as frequent or more frequent than the second most attested order. Otherwise, we assign the label “nondominant order”. In this case, once we have determined the different orders, we must consider only the SOV or OSV languages, which are the two that place subject and direct object dependencies before the verb in their basic dominant order.

Since in the languages we have available for analysis, we have not found any whose basic dominant order is OSV (since it is practically a marginal order), the only languages we will review are the SOV languages, that is Akkadian, Akuntsu, Amharic, Basque, Buryat, Hindi, Kazakh, Korean, Kurmanji, Marathi, Persian, Sanskrit, Tamil, Telugu, Turkish, Turkish-German, Uyghur, Xibe, Yakut.

Once we know the languages we will review, we must turn to the resource:
{http://tables.grew.fr/?data=ud_feats/FEATS}.

There we can check quickly and visually, line by line, the treebanks with the feature “case” and the type of case they have per feature. You can sort by frequency, move columns, or whatever the researcher wishes for a more comfortable reading.

One of our difficulties is considering what it means for a language to have a “case system”. Greenberg is not precise in this respect. Does the fact of having two cases already make it a system? If a language uses cases only in a specific part of speech, would we also consider it a case system?

We have not found any answer in the literature to address the question clearly. Therefore, we try to be maximalist with our proposal, in order to be able to represent the maximum number of possible elements. We set the usual filtering filtering in each case to be considered existing in a language. We believe that if a treebank cannot offer sufficient reliability of this minimum number of occurrences, it should not be considered since the proportionally relevant results may be misleading.

We propose simplifying the expression “case system” with “cases”, something more transparent. Therefore, we understand that establishing a distinction between two different cases is already relevant as opposed to languages that do not have such a feature.

21.41.2 Formalization Proposed for Grew-Match

On the other hand, we can calculate the presence of cases in the languages through Grew-Match without using tables and through formulations similar to those we have been proposing in this work. On the one hand, we can check the presence of cases in general in the languages using a general and straightforward formulation such as (21.545).

$$\begin{array}{l} \%41 - CASE \\ pattern\{CASE[Case]\} \end{array} \quad (21.545)$$

(21.545) captures any case occurrence.

In this, we have created a pattern that looks for any part of speech with the case feature active with any value.

However, since we want to know more precisely the behavior of cases, we also propose an individual analysis of each of the features within the case that languages may exhibit to establish different claims. For example, we will try to check if there is any kind of latent implicational hierarchy. To do this, we must take advantage of the previous pattern created to capture occurrences of cases in general and add just after the symbol = and the corresponding abbreviation with the value of the Universal feature label proposed by Universal Dependencies. In this case, the annotation scheme foresees the possibility of 38 different case types. We must formalize each one as in (21.546) - (21.582).

$$\begin{array}{l} \%41 - CASE - abs \\ pattern\{ABS[Case = Abs]\} \end{array} \quad (21.546)$$

(21.546) captures absolute cases.

$$\begin{array}{l} \%41 - CASE - acc \\ pattern\{ACC[Case = Acc]\} \end{array} \quad (21.547)$$

(21.547) captures accusative cases.

$$\begin{array}{l} \%41 - CASE - erg \\ pattern\{ERG[Case = Erg]\} \end{array} \quad (21.548)$$

(21.548) captures ergative cases.

$$\begin{array}{l} \%41 - CASE - Nom \\ pattern\{NOM[Case = Nom]\} \end{array} \quad (21.549)$$

(21.549) captures nominative cases.

$\%41 - CASE - abe$ $pattern\{ABE[Case = Abe]\}$	(21.550)

(21.550) captures abessive cases.

$\%41 - CASE - ben$ $pattern\{BEN[Case = Ben]\}$	(21.551)

(21.551) captures benefactive cases.

$\%41 - CASE - cau$ $pattern\{CAU[Case = Cau]\}$	(21.552)

(21.552) captures causative cases.

$\%41 - CASE - cmp$ $pattern\{CMP[Case = Cmp]\}$	(21.553)

(21.553) captures comparative cases.

$\%41 - CASE - cns$ $pattern\{CNS[Case = Cns]\}$	(21.554)

(21.554) captures considerative cases.

$\%41 - CASE - com$ $pattern\{COM[Case = Com]\}$	(21.555)

(21.555) captures comitative cases.

$\%41 - CASE - dat$ $pattern\{DAT[Case = Dat]\}$	(21.556)

(21.556) captures dative cases.

$\%41 - CASE - dis$ $pattern\{DIS[Case = Dis]\}$	(21.557)

(21.557) captures distributive cases.

$\%41 - CASE - equ$ $pattern\{EQU[Case = Equ]\}$	(21.558)

(21.558) captures equative cases.

$\%41 - CASE - gen$ $pattern\{GEN[Case = Gen]\}$	(21.559)

(21.559) captures genitive cases.

$\%41 - CASE - ins$ $pattern\{INS[Case = Ins]\}$	(21.560)

(21.560) captures instrumental cases.

$\%41 - CASE - par$ $pattern\{PAR[Case = Par]\}$	(21.561)

(21.561) captures partitive cases.

$\%41 - CASE - tem$ $pattern\{TEM[Case = Tem]\}$	(21.562)

(21.562) captures temporal cases.

$\%41 - CASE - tra$ $pattern\{TRA[Case = Tra]\}$	(21.563)

(21.563) captures translative cases.

$\%41 - CASE - voc$ $pattern\{VOC[Case = Voc]\}$	(21.564)

(21.564) captures vocative cases.

$\%41 - CASE - abl$ $pattern\{ABL[Case = Abl]\}$	(21.565)

(21.565) captures ablative cases.

$\%41 - CASE - add$ $pattern\{ADD[Case = Add]\}$	(21.566)

(21.566) captures additive cases.

$\%41 - CASE - ade$ $pattern\{ADE[Case = Ade]\}$	(21.567)

(21.567) captures adessive cases.

$\%41 - CASE - all$ $pattern\{ALL[Case = All]\}$	(21.568)

(21.568) captures allative cases.

$\%41 - CASE - del$ $pattern\{DEL[Case = Del]\}$	(21.569)

(21.569) captures delative cases.

$\%41 - CASE - ela$ $pattern\{ELA[Case = Ela]\}$	(21.570)

(21.570) captures elative cases.

$\%41 - CASE - ess$ $pattern\{ESS[Case = Ess]\}$	(21.571)

(21.571) captures essive cases.

$\%41 - CASE - ill$ $pattern\{ILL[Case = Ill]\}$	(21.572)

(21.572) captures illative cases.

$\%41 - CASE - ine$ $pattern\{INE[Case = Ine]\}$	(21.573)

(21.573) captures inessive cases.

$\%41 - CASE - lat$ $pattern\{LAT[Case = Lat]\}$	(21.574)

(21.574) captures lative cases.

$\%41 - CASE - loc$ $pattern\{LOC[Case = Loc]\}$	(21.575)

(21.575) captures locative cases.

$\%41 - CASE - per$ $pattern\{PER[Case = Per]\}$	(21.576)

(21.576) captures perlative cases.

$\%41 - CASE - sbe$ $pattern\{SBE[Case = Sbe]\}$	(21.577)

(21.577) captures subrelative cases.

$\%41 - CASE - sbl$ $pattern\{SBL[Case = Sbl]\}$	(21.578)

(21.578) captures sublative cases.

$\%41 - CASE - spl$ $pattern\{SPL[Case = Spl]\}$	(21.579)

(21.579) captures superlative cases.

$\%41 - CASE - sub$ $pattern\{SUB[Case = Sub]\}$	(21.580)

(21.580) captures subessive cases.

$\%41 - CASE - sup$ $pattern\{SUP[Case = Sup]\}$	(21.581)

(21.581) captures superessive cases.

$$\boxed{\begin{array}{l} \%41 - CASE - ter \\ pattern\{TER[Case = Ter]\} \end{array}} \quad (21.582)$$

(21.582) captures terminative cases.

21.41.3 Trichotomous Expression Based on Grew-Match Data

This interesting Universal about the performance of languages with dominant nominal SOV and OSV order exhibits 89% compliance in our set. However, in this case, we must also point out a small bias detected in Telugu. This language is a perfect example of what Greenberg comments: SOV-type language with a clear case system. However, through the data obtained in UD, it can be inferred that it is a language that does not meet such a Universal. This is because the Treebank annotators of this language have chosen to include all the information regarding the most elementary syntax (dependencies) and the part of speech. However, absolutely no features have been included. This means we cannot capture the fact that it has a case system and causes this problem. Therefore, given the importance of the tool and the data it allows us to capture (demonstrated in this thesis), we recommend a rigorous review of the annotation carried out, which will allow us to obtain even more accurate results.

In fact, in cases such as the one described above, we believe the approach proposed in the figure 21.74 is advantageous and valuable.

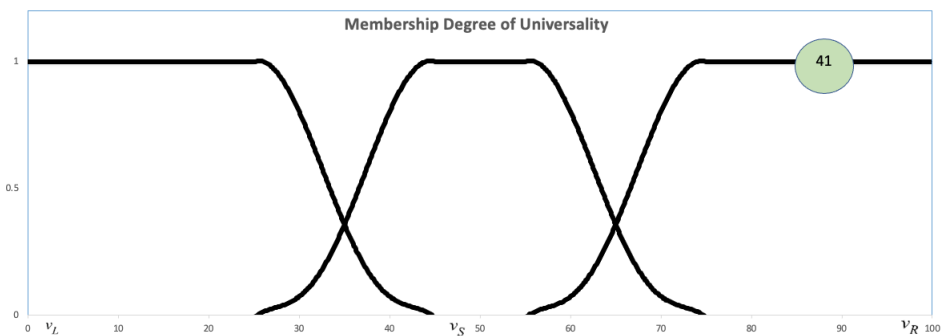


Figure 21.74: Trichotomous Expression for Universal 41

We can see how this 89% translates into a prototypical High-Universal on such a scale. If our approach were quantitative and we were to point out that 89% of SOV or OSV languages have a case system, such data would be false and easily disproved in future updates. Therefore, it would be outdated and not very useful. However, we consider this Universal a prototypical High-Universal

in the model presented. The changes that may occur in the future (correction of these small errors, addition of more languages, addition of more corpora in existing languages), will slightly modify the location of the Universal on the scale, without this change being significant. This assumes, therefore, that nothing changes for the classification and use of the Universal. We can summarise our findings as in table 21.39.

Summary of characteristics of the Universal 41 analysis:

1. Inconsistencies have been detected in the annotation of languages such as Telugu.
2. Greenberg’s original Universal is fulfilled in 89%.
3. Greenberg’s original Universal is a High-Universal.

Table 21.39: Summary of Characteristics of the Universal 41 Analysis

21.42 Universal 42

UNIVERSAL 42: *All languages have pronominal categories involving at least three persons and two numbers.*

21.42.1 Extracting the Information in the Treebank Hub Page

This Universal can be checked via the treebank hub page since we have just to confirm the quantity of variables in a Universal feature in pronouns. For this, we go to the treebank hub page of every corpus, and then we go to “POS Tags> PRON”. Then, we go to the morphology chapter and look for the occurrences with feature-value pairs. We should see that there are at least three different occurrences regarding the person and at least two different occurrences regarding nouns. For example, Spanish (AnCora) shows three persons (Person=1, Person=2, Person=3) and two numbers (Number=Plur, Number=Sing).

21.42.2 Formalization Proposed for Grew-Match

Checking this Universal via Grew-Match requires only a few steps. We must propose different patterns about a single element (without dependency), which always has the pronoun part of speech. Subsequently, we must outline its features in each pattern. We start by creating different patterns

with the different person possibilities (0,1,2,3,4), and then we check the number of possibilities we have already mentioned above. For the Universal to be true, we must find occurrences in 3 different patterns referring to persons and two different patterns referring to numbers. We formalize as in (21.583) - (21.598).

$$\begin{array}{l} \%42 - PRON - P0 \\ pattern\{PRON[upos = PRON, Person = 0]\} \end{array} \quad (21.583)$$

(21.583) captures pronouns with zero person value.

$$\begin{array}{l} \%42 - PRON - P1 \\ pattern\{PRON[upos = PRON, Person = 1]\} \end{array} \quad (21.584)$$

(21.584) captures pronouns with first-person value.

$$\begin{array}{l} \%42 - PRON - P2 \\ pattern\{PRON[upos = PRON, Person = 2]\} \end{array} \quad (21.585)$$

(21.585) captures pronouns with second person value.

$$\begin{array}{l} \%42, PRON - P3 \\ pattern\{PRON[upos = PRON, Person = 3]\} \end{array} \quad (21.586)$$

(21.586) captures pronouns with third person value.

$$\begin{array}{l} \%42, PRON - P4 \\ pattern\{PRON[upos = PRON, Person = 4]\} \end{array} \quad (21.587)$$

(21.587) captures pronouns with fourth person value.

$$\begin{array}{l} \%42 - PRON - Nsing \\ pattern\{PRON[upos = PRON, Number = Sing]\} \end{array} \quad (21.588)$$

(21.588) captures pronouns with singular numbers.

$$\begin{array}{l} \%42 - PRON - Nplur \\ pattern\{PRON[upos = PRON, Number = Plur]\} \end{array} \quad (21.589)$$

(21.589) captures pronouns with plural number.

$$\begin{array}{l} \%42 - PRON - Ndual \\ pattern\{PRON[upos = PRON, Number = Dual]\} \end{array} \quad (21.590)$$

(21.590) captures pronouns with dual number.

$$\begin{array}{l} \%42 - PRON - Ntri \\ pattern\{PRON[upos = PRON, Number = Tri]\} \end{array} \quad (21.591)$$

(21.591) captures pronouns with trial number.

$$\begin{array}{l} \%42, PRON - Npauc \\ pattern\{PRON[upos = PRON, Number = Pauc]\} \end{array} \quad (21.592)$$

(21.592) captures pronouns with a paucal number.

$$\begin{array}{l} \%42 - PRON - Ninv \\ pattern\{PRON[upos = PRON, Number = Inv]\} \end{array} \quad (21.593)$$

(21.593) captures pronouns with inverse number.

$$\begin{array}{l} \%42 - PRON - Ncoll \\ pattern\{PRON[upos = PRON, Number = Coll]\} << \end{array} \quad (21.594)$$

(21.594) captures pronouns with collective number.

$$\begin{array}{l} \%42 - PRON - Ncount \\ pattern\{PRON[upos = PRON, Number = Count]\} \end{array} \quad (21.595)$$

(21.595) captures pronouns with countable number.

$$\begin{array}{l} \%42 - PRON - NGrpa \\ pattern\{PRON[upos = PRON, Number = Grpa]\} \end{array} \quad (21.596)$$

(21.596) captures pronouns with greater paucal number.

$$\begin{array}{l} \%42 - PRON - NGrpl \\ pattern\{PRON[upos = PRON, Number = Grpl]\} \end{array} \quad (21.597)$$

(21.597) captures pronouns with greater plural number.

$$\begin{array}{l} \%42 - PRON - NPtan \\ pattern\{PRON[upos = PRON, Number = Ptan]\} \end{array} \quad (21.598)$$

(21.598) captures pronouns with plurale tantum number.

21.42.3 Trichotomous Expression Based on Grew-Match Data

This other Universal deals with a characteristic, a priori, present in all the world’s languages, regardless of their type. We must state in this case that we have also detected some punctual oversight in some languages, but this is not a widespread situation that would hinder the consideration of the Universal. However, we point this out so that these data, which affect some languages individually, can be corrected.

The problems detected can be seen in treebanks such as Dutch, where the error seems to be generated by an annotator’s laziness. We come across the personal pronoun “we”, equivalent to the English “we”. The information given for this element is “[upos=PRON, lemma=we, Case=Nom, Person=1, PronType=Prs]”. No information concerning the number appears when there is a clear opposition between the “we”, the plural form, and “ik”, the singular form.

Figure 21.75 shows the trichotomous expression representing this Universal.

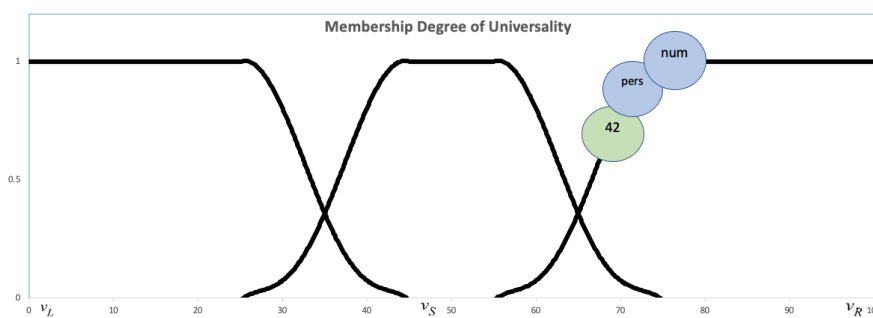


Figure 21.75: Trichotomous Expression for Universal 42

The analyzed items are placed in the High-Universal set, although they do not present this set’s maximum values of representativeness. In other words, this claim may be true for most languages and can be used as a general rule. However, this High-Universal is not a clear example of a law shared by all languages in the world. Therefore, we expect to see future results that can be compared with those provided here, especially once the annotation aspect noted is revised. This may slightly modify the location of the point on the scale, although we believe not so significantly

that it leaves the set it is in, one of the advantages we believe this approach has. We can summarise our findings as in table 21.40.

Summary of characteristics of the Universal 42 analysis: 1. The overall fulfillment of the original Universal is 70%. 2. This comes from an acceptance of 73% in the case of exhibiting three persons in the pronouns. 3. An acceptance of 75% is observed in exhibiting two different numbers in the pronouns. 4. The original Universal is a High-Universal. Also, the two parts. However, it should be carefully reviewed in future research, as it does not have the maximum value of belonging to the High set.

Table 21.40: Summary of Characteristics of the Universal 42 Analysis

21.43 Universal 43

UNIVERSAL 43: *If a language has gender categories in the noun, it has gender categories in the pronoun.*

21.43.1 Extracting the Information in the Treebank Hub Page

This Universal can be checked via treebank hub page, since we have just to confirm the presence of one feature in nouns and pronouns. For this, we go to the treebank hub page of every corpus, and then we go to “POS Tags> NOUN” and “POS Tags>PRON”. Then, we go to the morphology chapter and check if the “gender” feature is active in them. For example, in Spanish (AnCora), we find “gender” in nouns and pronouns.

21.43.2 Formalization Proposed for Grew-Match

If we want to check this Universal in Grew-Match, we must simply formalize an element with a given part of speech (noun in one pattern and pronoun in another) and ask for any gender feature to be active. However, in this case, we also want to check another part of speech to have a complete view of the application of the gender feature to the part of speech of languages. Therefore, we

propose formalizations in all possible ones. This makes the formalizations we propose (21.599) - (21.611).

$\%43 - NOUN$ $pattern\{NOUN[upos = NOUN, Gender]\}$	(21.599)

(21.599) captures nouns with any gender value.

$\%43 - PRON$ $pattern\{PRON[upos = PRON, Gender]\}$	(21.600)

(21.600) captures pronouns with any gender value.

$\%43 - ADJ$ $pattern\{ADJ[upos = ADJ, Gender]\}$	(21.601)

(21.601) captures adjectives with any gender value.

$\%43 - ADV$ $pattern\{ADV[upos = ADV, Gender]\}$	(21.602)

(21.602) captures adverbs with any gender value.

$\%43 - INTJ$ $pattern\{INTJ[upos = INTJ, Gender]\}$	(21.603)

(21.603) captures interjections with any gender value.

$\%43 - VERB$ $pattern\{VERB[upos = VERB, Gender]\}$	(21.604)

(21.604) captures verbs with any gender value.

$\%43 - ADP$ $pattern\{ADP[upos = ADP, Gender]\}$	(21.605)

(21.605) captures adpositions with any gender value.

$\%43 - AUX$ $pattern\{AUX[upos = AUX, Gender]\}$	(21.606)

(21.606) captures auxiliaries with any gender value.

$\%43 - CONJ$ $pattern\{CONJ[upos = CONJ, Gender]\}$	(21.607)

(21.607) captures coordinative conjunctions with any gender value.

$\%43 - DET$ $pattern\{DET[upos = DET, Gender]\}$	(21.608)

(21.608) captures determinants with any gender value.

$\%43 - NUM$ $pattern\{NUM[upos = NUM, Gender]\}$	(21.609)

(21.609) captures numerals with any gender value.

$\%43 - PART$ $pattern\{PART[upos = PART, Gender]\}$	(21.610)

(21.610) captures particles with any gender value.

$\%43 - SUBJ$ $pattern\{SUBJ[upos = SUBJ, Gender]\}$	(21.611)

(21.611) captures subordinate conjunctions with any gender value.

21.43.3 Trichotomous Expression Based on Grew-Match Data

Again, the results obtained in this Universal should be reviewed and refined in the future, due to some errors found in the annotation of gender in some words. Usually, we have detected errors in the case of automatic procedures, which assign gender to words that do not have it, such as “everybody” or “tax” in English, which appear as feminine.

However, we believe the results we obtain represent the reality of Universal compliance for two reasons. First, the statement that English does not exhibit gender in nouns (but does in pronouns)

or that English exhibits gender in both would be valid and correct contributions to the Universal, being placed in different methods of fulfillment but not representing an exception.

Secondly, a fine-grained and specific approach, such as the one we have carried out, allows us to reach specific examples that allow us to specify and profile much more precise statements, such as those made in more general approaches. We believe that English does express gender differences in nouns, but it does not do so by inflection and not in most words. For example, we find this difference marked in words such as King/Queen; there, we see the gender marking of nouns as reasonable. Therefore, these residual, marginal cases, which may go unnoticed, may be relevant.

The Universal in its original version is fulfilled in 95% of the cases. If we replace “nouns” with “adjectives” in the claim, we obtain a 100% success rate. In the case of proper nouns, we see that it is fulfilled 88% of the time. Regarding verbs, it is true 98% of the time and 97% of the time for auxiliaries. “If a language has gender markers in determiners, it has gender markers in pronouns” would be true in 98% of the cases, while it is true in 100% of the cases for numerals.

The graphical representation of such data with our model appears in figure 21.76.

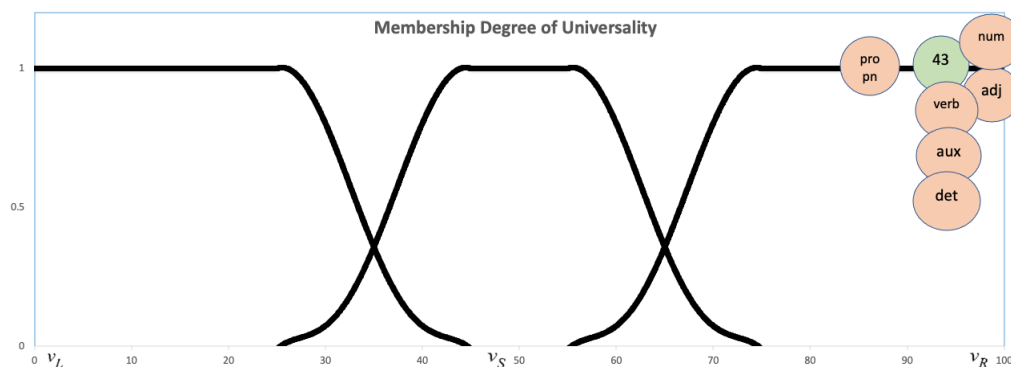


Figure 21.76: Trichotomous Expression for Universal 43

All the previously presented checks correspond to a High-Universal at the highest possible representativeness point. The Universal proposed by Greenberg is shown to be accurate, and therefore, there is a clear preference for marking the gender of pronouns in the first place (if the language marks the gender). In the second place, we find the rest of the parts of speech with very similar values, which allow us to formulate an alternative and extended Universal, with greater coverage and representation, around the primacy of the presence of gender in pronouns as a prior step to such use in other varied parts of speech that had not been contemplated. We can summarise our findings as in table 21.41.

Summary of characteristics of the Universal 43 analysis: 1. In the process of verifying this Universal, we have also detected annotation problems, even in English, about gender. 2. We can capture non-prototypical occurrences regarding gender characterization in some languages in which it exists in very specific cases. 3. Greenberg’s original Universal obtains a result of 95%. 4. We propose the verification of a Universal focused on adjectives and numerals instead of nouns, and we obtain 100% compliance. 5. In the case of the review of other parts of speech, such as proper nouns (88%), verbs (98%), auxiliaries (97%), and determiners (98%), we also obtain valid and interesting results. 6. Both the original Greenberg Universal and all the extra checks we have carried out are High-Universals.

Table 21.41: Summary of Characteristics of the Universal 43 Analysis

21.44 Universal 44

UNIVERSAL 44: *If a language has gender distinctions in the first person, it always has gender distinctions in the second or third person or both.*

21.44.1 Formalization Proposed for Grew-Match

Universal 44 claims that if in our pattern there is *Person* = 1 plus *Gender* = *Fem*, for example, it means that we are going to find also *Gender* = *Fem* in *Person* = 2 or *Person* = 3 or both of them. It is interesting, thus, not only to check if this Universal is confirmed, but also to see which of the two persons is more prone to these agreements or if there is any further invariance typologically. Then, instead of searching the second and third person together, we opt to search them separately. Confirming one of the two patterns will be enough to consider the Universal valid.

To further extend the application knowledge of this Universal, we can also check the behavior of different parts of speech, such as auxiliaries. Furthermore, according to the UD annotation scheme, we extend the analysis to all possible persons. In addition to the first, second, and third

proposed by Greenberg, we also analyze what happens with the persons labeled 0 or 4 in UD. These persons, explained in the introductory annotation scheme to this section, correspond to much less widespread and Universal cases. Therefore, we expect the number of occurrences found to be much lower.

In addition to checking that the verb has any active gender feature, we performed a specific check with the feminine gender as an example. We believe that the selection of this gender helps to avoid possible generation errors due to the lack of distinction between masculine and non-gender features. Although verbs are mentioned in the commentary before the Greenberg Universal, no part of speech is mentioned in the Universal itself. Therefore, we also decided to apply formalization to any part of speech. As mentioned above, we simply have to combine the different person possibilities we have (0,1,2,3,4) with the presence of any gender label. We also see what results we get when the gender label is feminine, changing “*Gender*” for “*Gender = Fem*”. We do this on the verb part of speech, the auxiliary part, and any part of speech. We propose the patterns (21.612) - (21.641) for this.

$$\begin{array}{l} \%44 - 0 \\ pattern\{V[upos = VERB, Person = 0, Gender = Fem]\} \end{array} \quad (21.612)$$

(21.612) captures verbs with zero person and feminine gender.

$$\begin{array}{l} \%44 - 1 \\ pattern\{V[upos = VERB, Person = 1, Gender = Fem]\} \end{array} \quad (21.613)$$

(21.613) captures verbs with first person and feminine gender.

$$\begin{array}{l} \%44 - 2 \\ pattern\{V[upos = VERB, Person = 2, Gender = Fem]\} \end{array} \quad (21.614)$$

(21.614) captures verbs with a second person and feminine gender.

$$\begin{array}{l} \%44 - 3 \\ pattern\{V[upos = VERB, Person = 3, Gender = Fem]\} \end{array} \quad (21.615)$$

(21.615) captures verbs with a third person and feminine gender.

$\%44 - 4$ $pattern\{V[upos = VERB, Person = 4, Gender = Fem]\}$	(21.616)

(21.616) captures verbs with the fourth person and feminine gender.

$\%44 - 0Aux$ $pattern\{AUX[upos = AUX, Person = 0, Gender = Fem]\}$	(21.617)

(21.617) captures auxiliaries with zero person and feminine gender.

$\%44 - 1Aux$ $pattern\{AUX[upos = AUX, Person = 1, Gender = Fem]\}$	(21.618)

(21.618) captures auxiliaries with first person and feminine gender.

$\%44 - 2Aux$ $pattern\{AUX[upos = AUX, Person = 2, Gender = Fem]\}$	(21.619)

(21.619) captures auxiliaries with a second person and feminine gender.

$\%44 - 3Aux$ $pattern\{AUX[upos = AUX, Person = 3, Gender = Fem]\}$	(21.620)

(21.620) captures auxiliaries with a third person and feminine gender.

$\%44 - 4Aux$ $pattern\{AUX[upos = AUX, Person = 4, Gender = Fem]\}$	(21.621)

(21.621) captures auxiliaries with the fourth person and feminine gender.

$\%44 - 0 - feat$ $pattern\{V[upos = VERB, Person = 0, Gender]\}$	(21.622)

(21.622) captures verbs with zero person and any gender.

$\%44 - 1 - feat$ $pattern\{V[upos = VERB, Person = 1, Gender]\}$	(21.623)

(21.623) captures verbs in the first person and any gender.

$$\begin{array}{l} \%44 - 2 - feat \\ pattern\{V[upos = VERB, Person = 2, Gender]\} \end{array} \quad (21.624)$$

(21.624) captures verbs with the second person and any gender.

$$\begin{array}{l} \%44 - 3 - feat \\ pattern\{V[upos = VERB, Person = 3, Gender]\} \end{array} \quad (21.625)$$

(21.625) captures verbs in the third person and any gender.

$$\begin{array}{l} \%44 - 4 - feat \\ pattern\{V[upos = VERB, Person = 4, Gender]\} \end{array} \quad (21.626)$$

(21.626) captures verbs with the fourth person and any gender.

$$\begin{array}{l} \%44 - 0Aux - feat \\ pattern\{AUX[upos = AUX, Person = 0, Gender]\} \end{array} \quad (21.627)$$

(21.627) captures auxiliaries with zero person and any gender.

$$\begin{array}{l} \%44 - 1Aux - feat \\ pattern\{AUX[upos = AUX, Person = 1, Gender]\} \end{array} \quad (21.628)$$

(21.628) captures auxiliaries with first person and any gender.

$$\begin{array}{l} \%44 - 2Aux - feat \\ pattern\{AUX[upos = AUX, Person = 2, Gender]\} \end{array} \quad (21.629)$$

(21.629) captures auxiliaries with the second person and any gender.

$$\begin{array}{l} \%44 - 3Aux - feat \\ pattern\{AUX[upos = AUX, Person = 3, Gender]\} \end{array} \quad (21.630)$$

(21.630) captures auxiliaries with the third person and any gender.

$$\begin{array}{l} \%44 - 4Aux - feat \\ pattern\{AUX[upos = AUX, Person = 4, Gender]\} \end{array} \quad (21.631)$$

(21.631) captures auxiliaries with the fourth person and any gender.

$\%44 - 0Any$ $pattern\{ANY[Person = 0, Gender = Fem]\}$	(21.632)

(21.631) captures any part of speech with zero person and feminine gender.

$\%44 - 1Any$ $pattern\{ANY[Person = 1, Gender = Fem]\}$	(21.633)

(21.633) captures any part of speech with first person and feminine gender.

$\%44 - 2Any$ $pattern\{ANY[Person = 2, Gender = Fem]\}$	(21.634)

(21.634) captures any part of speech with second person and feminine gender.

$\%44 - 3Any$ $pattern\{ANY[Person = 3, Gender = Fem]\}$	(21.635)

(21.635) captures any part of speech with third person and feminine gender.

$\%44 - 4Any$ $pattern\{ANY[Person = 4, Gender = Fem]\}$	(21.636)

(21.636) captures any part of speech with the fourth person and feminine gender.

$\%44 - 0 - feat - any$ $pattern\{ANY[Person = 0, Gender]\}$	(21.637)

(21.637) captures any part of speech with zero person and any gender.

$\%44 - 1 - feat - any$ $pattern\{ANY[Person = 1, Gender]\}$	(21.638)

(21.638) captures any part of speech in the first person and any gender.

$\%44 - 2 - feat - any$ $pattern\{ANY[Person = 2, Gender]\}$	(21.639)

(21.639) captures any part of speech with the second person and any gender.

$\%44 - 3 - \textit{feat} - \textit{any}$ $\textit{pattern}\{ANY[Person = 3, Gender]\}$	(21.640)

(21.640) captures any part of speech with the third person and any gender.

$\%44 - 4 - \textit{feat} - \textit{any}$ $\textit{pattern}\{ANY[Person = 4, Gender]\}$	(21.641)

(21.641) captures any part of speech with the fourth person and any gender.

21.44.2 Trichotomous Expression Based on Grew-Match Data

The gender distinctions in the different persons show very interesting results. We capture interesting occurrences for study in both verbs and possessives.

In this case, linguistic knowledge or intuition may suggest this is true for linguists. However, it is convenient to base the claims on objective and verifiable data and, in addition, to be able to review the data in multiple languages (far from the bias of each researcher) and in authentic texts.

1. In this case, we can see how the compliance with the Universal is 100%.
2. In addition, we have been able to review what percentage of compliance is given in each of the possibilities enunciated by Greenberg. 100% of the languages that have a distinction in the first person also establish this distinction in the third person.
3. 90% of the languages that establish a gender distinction in the first person also establish this distinction in the second person.
4. Furthermore, 90% of the languages that establish gender distinction in the first person establish gender distinction in the second and third person.

Therefore, it seems that a possible implicational hierarchy similar to:

1. No gender distinction in any person.
2. Gender distinction in the third person.
3. Gender distinction in the second person.
4. Gender distinction in the first person.

It would also be interesting to review the role of other less extended persons, such as the fourth person, and their placement in this drawn hierarchy.

The percentages previously presented are placed on a trichotomous scale as in figure 21.77.

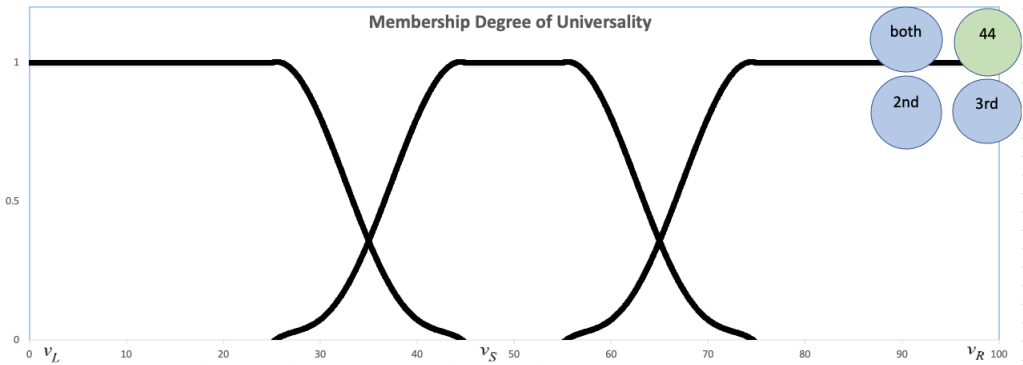


Figure 21.77: Trichotomous Expression for Universal 44

In a clear way, Greenberg’s Universal is a high Universal with the maximum value of prototypicality, as well as the different parts that he points out as possible conditions that allow this Universal to be fulfilled. It is, therefore, a Universal well formulated by Greenberg, although its use in Typology is more restricted. We can summarise our findings as in table 21.42.

Summary of characteristics of the Universal 44 analysis:
1. The original Greenberg Universal has 100% compliance.
2. The third person is the one that presents 100% gender distinction if it also appears in the first person.
3. The second person presents a 90% gender distinction if it also appears in the first person.
4. The gender distinction always exists in the third person (if it exists). In many cases also in the second person. Lastly is the first person.
5. The original Greenberg Universal and the more precise checks we have carried out are all High-Universals.

Table 21.42: Summary of Characteristics of the Universal 44 Analysis

21.45 Universal 45

UNIVERSAL 45: *If there are any gender distinctions in the plural of the pronoun, there are some gender distinctions in the singular.*

21.45.1 Formalization Proposed for Grew-Match

In this Universal, we must work with the part of speech of *PRON*. We must collect in our formalizations the possibility of any gender value appearing in singular forms and then in plural forms to be able to contrast the occurrence data. To this end, we propose (21.642) and (21.643).

$\%45 - PRON_{pl}$ $pattern\{PRON[upos = PRON, Gender, Number = Plur]\}$	(21.642)

(21.642) captures pronouns with any gender value and plural number.

$\%45 - PRON_{sg}$ $pattern\{PRON[upos = PRON, Gender, Number = Sing]\}$	(21.643)

(21.643) captures pronouns with any gender value and singular number.

However, we are interested in checking if there is any other part of speech that behaves in the same way as pronouns. Therefore, we apply this same distinction to other parts of speech, something that could allow us to extend the Universal. For example, we show this application to adjectives in (21.644) and (21.645). However, we have checked other parts of speech possibilities by changing the formalization to the corresponding part of speech.

$\%45 - ADJ_{pl}$ $pattern\{ADJ[upos = ADJ, Gender, Number = Plur]\}$	(21.644)

(21.644) captures adjectives with any gender value and plural number.

$\%45 - ADJ_{sg}$ $pattern\{ADJ[upos = ADJ, Gender, Number = Sing]\}$	(21.645)

(21.645) captures adjectives with any gender value and singular number.

21.45.2 Trichotomous Expression Based on Grew-Match Data

Once again, we are faced with a Universal that is not very typological and more general, which, coherently, offers 100% compliance. In addition, we wanted to check whether this rule applies to other parts of speech. In the case of adjectives, verbs, auxiliaries, and determiners, we found 100% compliance. This may lead us to formulate a new Universal in which the reference to the part of speech “pronoun” is eliminated and, therefore, to extend further this characteristic shared by the world’s languages.

This arrangement we have discussed would be represented graphically as in figure 21.78.

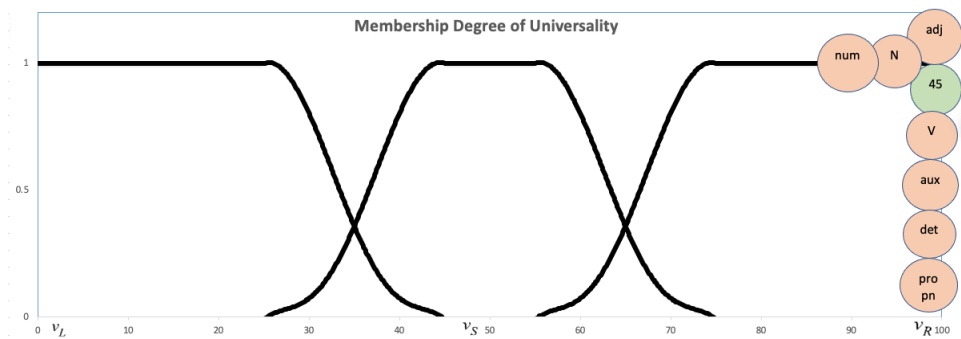


Figure 21.78: Trichotomous Expression for Universal 45

All the elements tested are High-Universal and representative at their maximum value. The different possible formulations of this High-Universal (and its original version) mark, in this case, a clear behavior shared by all the world’s languages, regardless of their type. We can summarise our findings as in table 21.43.

Summary of characteristics of the Universal 45 analysis:
1. The original Greenberg Universal has a 100% compliance rate.
2. Adjectives, verbs, auxiliaries, determiners, and proper nouns also exhibit 100 percent compliance with the same conditional structure as the Greenberg Universal. In the case of nouns and numerals, the percentage is close to 95% and 90%, respectively.
3. Both the original Greenberg Universal and the rest of the part of speech we have reviewed are High-Universals.

Table 21.43: Summary of Characteristics of the Universal 45 Analysis

UNIVERSITAT ROVIRA I VIRGILI

GLU - A GRADIENT MODEL FOR LANGUAGE UNIVERSALS. A COMPUTATIONAL REVISITATION OF GREENBERG UNIVERSALS

Antoni Brosa Rodríguez

22

General Results

Contents

22.1	Quantitative Review of All Greenberg Universals	577
22.2	Quantitative Review of the Proposed High-Universals	578

22.1 Quantitative Review of All Greenberg Universals

Once we have carried out this individual review of each of Greenberg's Universals and presented the corresponding data, we believe it is interesting to see these results as a whole. Our GLU Model allows us to locate all these reviewed elements to appreciate their relationship with each other and their arrangement. As there are so many Universals, in some cases, there is overlap in the same position. In this case, we have chosen to place these same Universals in a different position on the vertical axis, respecting their location on the horizontal axis. We believe the loss of information is minimal because it is rescued by looking at the prototypicality of the Universal with which it shares an axis. For example, Universal 36 is a prototypical High-Universal (1), comparable to Universal 3, which is the one located where it corresponds in the figure 22.1.

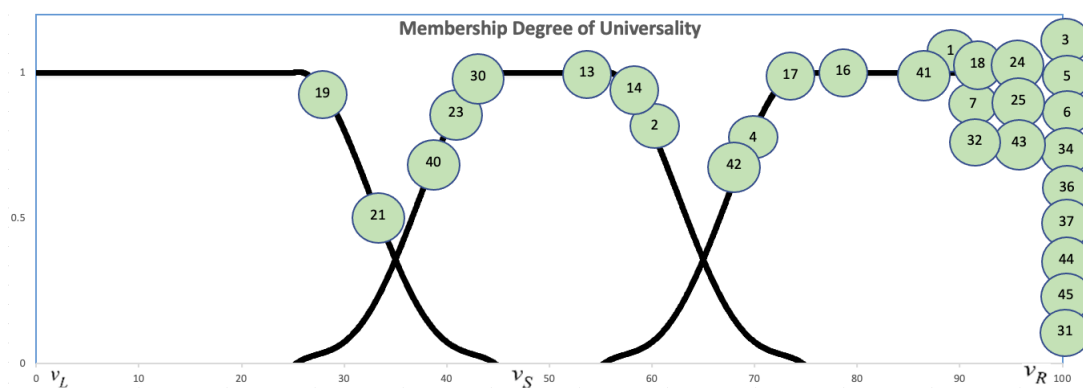


Figure 22.1: Trichotomous Expression for Greenberg's Original Universals

In this image, different data can be interpreted, corresponding to the state of the Universals in general. Of the 45 claims made by Greenberg, initially, in the formalization, we have discarded 11 that are impossible to review with the available information, given the limitations of Universal Dependencies, which do not even have an annotation system for these claims since they do not exist. Of these 34 claims that we have been able to formalize, we have found three Universals in which the errors appear in the data search due to problems in the annotation. In addition, we do not show Greenberg's own two Universals in which it has not been possible to carry out an exact representation of their formulation but a more general approximation. This means that, in total, we offer the exact representation, based on real texts of the languages available in UD, of 29 Universals formulated by Greenberg.

Of those 29 Universals that we offer revised, there are 21 of them that are High-Universal. This means that 72.5% of the claims proposed by Greenberg and which have been reliably checked are High-Universals and, therefore, make an interesting contribution to the knowledge of the functioning of the world's languages. In other words, 27.5% of the Universals formulated by Greenberg are not helpful for typological knowledge of the world's languages, according to our data.

22.2 Quantitative Review of the Proposed High-Universals

After seeing the general behavior of Greenberg's original claims, we would like to show our proposal once all the data have been checked. In the case of respecting Greenberg's proposal, the color green is maintained. In the case of pointing out our modification as a more adequate formulation to represent a phenomenon, it appears in orange. The results can be seen in figure 22.2.

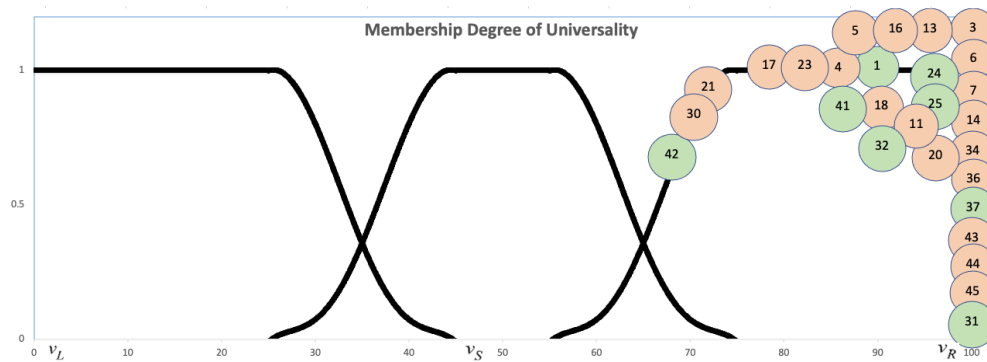


Figure 22.2: Proposal of the Trichotomous Expression for Greenberg's Universals

We must remember that of the 45 possible Universals formulated by Greenberg, only 34 could be formalized due to the limitations of UD and the type of data included. Of that number, we must also remember that we cannot provide information on three Universals that have presented problems in the data collection step due to errors in the annotation of the Universal Dependencies resources.

We have previously commented that 8 Greenberg Universals failed to place in the High-Universal group. We have improved the formulation of 5 of them so that they become part of that group, thanks to different modifications in their formulation. This means that three Universals must be discarded as no alternative has been found.

In summary, in the graph, we can represent 28 different Universals. Of this number, eight are shown in green, therefore Greenberg's original formulation. The other 20 (orange color) are formulations based on Greenberg's Universal but with some improvement. In some cases, our version has been chosen in order to be able to change a Mid-Universal to a High-Universal. In other cases, however, the Greenberg Universal showed a similar value in High-Universal, but our formulation was considered more complete. In this case, we should understand "complete" as the possibility of extending the Universal to more elements, linguistic types, etc.

One of the most interesting proposals in Linguistic Typology, due to its predictive power and classificatory consequences (inter-linguistically and intralinguistically), is the implicational hierarchies. In Greenberg's 45 Universals, we do not formally find any broad hierarchy of this type, although some of his formulations may give rise to the continuation of some of them. It has been later, researchers carried out formulations of this type.

Thanks to our study, in which we tried to verify correlations beyond those merely formulated by Greenberg, we have verified possible hierarchies based on Universals 36 and 43.

In relation to Universal 36, we believe that there is a close relationship between all the morphological features, not only the two indicated by Greenberg. In most cases, we observe patterns in which the presence of one element appears whenever another appears, but not vice versa.

About Universal 43, we also believe that it is possible to hierarchize, in this case, the parts of speech affected. In other words, different implications can be concatenated, so if a given part of speech contains different gender categories, the other part of the Universal will have them.

We have not been able to develop these two implicational hierarchies because they are not the primary interest of this doctoral thesis. For example, there should be a correlation between the data obtained and the diachrony. However, we believe that the data we provide can be a good starting point to certify (or not) whether the information contained in Universals 36 and 43 can be formally proposed as an implicational hierarchy with the methods we have worked with.

Part VI

Conclusions

UNIVERSITAT ROVIRA I VIRGILI

GLU - A GRADIENT MODEL FOR LANGUAGE UNIVERSALS. A COMPUTATIONAL REVISITATION OF GREENBERG UNIVERSALS

Antoni Brosa Rodríguez

23

Conclusions

Contents

23.1	GLU: A Gradient Concept of Universality	584
23.2	Formalization of Greenberg Universals	589
23.3	Analysis of Greenberg Universals	595

We have started this doctoral thesis with the following questions:

1. Where is the cut-off point between something universal with exceptions and something that is not universal?
2. How can the presence or absence of proven exceptions to the Universals be formulated to be more fairly represented?
3. Is it possible to test or validate Greenberg Universals computationally?
4. Can we formulate new Universals from Greenberg ones?

Those questions have been used to set our five objectives:

1. To review the concept of Universals in Language.
2. To propose a new classification of Universals in Language, with special attention to the presence of exceptions.
3. To provide a formalization for Greenberg Universals.
4. To computationally check Greenberg Universals through real texts.

5. To revisit the formulation of some Greenberg Universals.

The fulfillment of these objectives and the response to them have generated a series of outputs that represent the contribution of this thesis to Linguistics. The outputs can be organized into three main blocks, each with two different aspects (the first is always more theoretical or review nature and the second of a more practical nature).

1. Proposal of the concept of Universality as gradient

- Review and compilation of the varied terminology of the Universals
- Theoretical proposal of a model

2. Formalization of Greenberg Universals

- Analysis of the advantages and limitations for a study of Greenberg Universals computationally
- Proposed formalization of Greenberg's Universals

3. Analysis of Greenberg Universals

- Testing Universals computationally with real texts
- Formulation of new Universals in language

In the following sections, we elaborate on each of the main contributions mentioned above.

23.1 GLU: A Gradient Concept of Universality

The doctoral thesis defends a different approach to the traditional one in Linguistic Typology concerning Language Universals. The object of study, the Universal, is understood as a linguistic rule present in the languages of the world. In our opinion, this rule has a certain degree of Universality, which can range from the lowest to the highest. Its degree depends on how many languages share the rule. Those rules with a high level of Universality are of interest to Linguistic Typology, as they allow for a deeper understanding of the languages of the world and their types.

This conception is opposed to the following classical guidelines for understanding universality in languages:

- Distinction between absolute and statistical Universals.
- Defence of Universals with exceptions.
- Lack of clarity in distinguishing Universals from non-Universals.
- Sharp boundaries to differentiate absolute Universals from statistical Universals, Universals from non-Universals, or other distinctions such as almost Universals (if established).

In our opinion, these characteristics condition the study of Universals in two main ways:

- Doubts about the existence and nature of Universals.
- Many different labels refer to Universals.

Doubts about Universals are fundamentally due to confusion about the term. Many authors use it for different purposes. In the case of Linguistic Typology, we have seen how the interest lies in explaining the characteristics shared by languages belonging to the same linguistic type. Once this is clarified, it is understood that there are clear correlations between linguistic types and other characteristics. The existence of typological Universals is, in fact, exemplified by precise data in this doctoral thesis. There may be multiple and punctual influences on some language and its concrete realization, but the functional explanation behind the Universal remains valid.

The existence of different labels is a symptom of the guidelines in the classical conception of Universals shown above, which has not yet been solved. We have not found an analysis of the labels linked to Universals in the literature and have decided to explore this issue.

23.1.1 Review and Compilation of the Universals Terminology

Approximately 60 different labels in the field of Universals studies are collected in the thesis. In many cases, these try to be euphemistic to the concept of Universal or try to qualify it. We have linked these proposals to the authors who proposed them and tried to understand the relationship between them. We have been able to gather them in three main groups:

1. **Labels relating to the Universal conception.** Difference between Universals formulated with Greenberg's methodology or formulated like Chomsky's. That is, it is related to the previous point, in which we commented that there are different purposes behind "Universal".

We find the opposition between surface and deep Universals, respectively. Within these two large groups, there are different labels used by different authors, which we have collected.

2. Labels relating to the structure with which the Universal is formulated. Difference between conditionally formulated Universals and Universals without conditional structure. Conditional Universals have a special relationship with typological Universals since they are the structures used to express the link between a linguistic type and one of its characteristics. Here, we find the distinction between implicational and unrestricted Universals, respectively. We have also collected a wide variety of labels.

3. Tags relating to the presence of exceptions in the Universal. Differentiate between Universals that have exceptions and those that do not. Universals with exceptions are those that are closely linked to typological Universals. The study of concrete data on the realizations of the different languages favors the appearance of some exceptions. We find statistical and absolute Universals, respectively. There is also a large number of alternative labels synonymous with these two that we have been able to note throughout the bibliography. In this case, we observe a particular interest in trying to change the label Universal or to qualify it as a euphemism in order to try to give little importance to the presence of exceptions.

This great variety of labels is not very positive. Their existence can hinder transparent knowledge in the field of Universals. A non-expert reader may not know whether they refer to the same element or not, which may make it difficult to access the information. In the case of expert readers, there are severe doubts about whether using one label or another has any implication about the nature of the Universal indicated. Despite the existence of all this terminology, the total or partial equivalence between the different existing terms is neither precise nor homogenized.

The annotated list that we offer with the different labels and the authors that support them is clear proof of the existing problem in the theoretical conception of Universals from the perspective of the presence or absence of exceptions. Many experts in Typology point out the need to naturalize the exceptions obtained. However, the framework used is still incompatible with such a view. On the one hand, the distinction between absolute and statistical Universals is maintained by most authors. On the other hand, categorizing a Universal as one or the other, or the differentiation between a rule that is or is not Universal, remains discrete and abrupt. This fact makes the labels used highly dependent on the sampling selected in each investigation, and, therefore, different samples use different labels for the same phenomenon. For example, if more precise internal distinctions are added, such as almost Universals, we find the same situation: there are sharp

boundaries between groups.

In the 21st century, we observed a greater tendency to use more euphemistic labels, such as “tendency”. In the cases observed, there was no explicit proposal advocating a change in the classification model. The only thing we observed was the exchange of one label for another.

In summary, we consider that many of the authors analyzed are aware of the limitations of the prevailing classification model. This discomfort can be seen in all the examples discussed in these paragraphs. However, we have not observed any analysis such as the one we have made where all these labels created by historical addition to the problem of Universality are collected systematically and referenced. Nor have we observed a classification proposal such as the one we offer for modeling Universals.

23.1.2 A Theoretical Proposal: GLU Model

This thesis offers typologists a different way of understanding Universality. There are several intended consequences of adopting the model we propose. Fundamentally, now, we can:

- **Merge statistical and absolute Universals under one group.** The model based on three major groups makes it possible to blur the abrupt borderline between statistical and absolute Universals. This is a claim of some typologists who advocated the primacy of statistical Universals, even if they did not offer any such concrete solution. In this way, the distinction becomes irrelevant in Typology, as is the case, *de facto*. Exceptions have been naturalized.
- **Avoid the continuous change of labels conditioned by languages.** The fact of having a statistical or absolute Universal, in many cases, depends on the sample of languages on which the study is based. This may result in the same Universal being considered differently by different authors. Any absolute Universal is liable to cease to be an absolute Universal at any time. With the proposal we are making, these uncertainties become secondary and no longer relevant since no such distinction is made.
- **Offer a concrete proposal for Universality.** The model assumes a concrete definition of Universal in which there are different degrees of Universality. The interesting rules are those that have a high degree of Universality. These are the ones we equate with the traditional concept of Universal. The model has a clearly marked group where these Universals or rules

are located. The model mathematically determines the size of this group: it is the same size we use to classify an object as large, medium, or small, for example.

- **Provide gradual boundaries between groups.** The end of the group we are interested in, i.e., the boundary between High-Universals and Mid-Universals, is presented with a gradual transition. This allows for less dependence on the concrete results obtained by a given sampling. The individual data may change slightly, but the Universal will not change groups easily. It is a sufficient change in consistency with the mathematical model that would allow such a group change. If it does occur, the same path must be followed to assert that we are in a High-Universal again. This allows for less hesitation and more consistency.
- **Unify the results of Universals from different samples.** The universe of languages used to formulate a Universal can vary enormously. Regarding implicational Universals, depending on the linguistic type to which the Universal refers, the number of languages it applies also varies concerning others. The graphical representation of the trichotomous model does not vary regardless of the total number of languages to which it applies, as reflected in the figures on the horizontal axis. This facilitates the comparison between Universals since what is relevant is the point on the scale where they are located.
- **Being less dependent on limited sampling.** A representation with sharp boundaries can excessively influence the modeling of Universals with concrete sampling. We want to carry out a study of Universals from Computational Linguistics. Data availability is not perfect. However, the proposed approach allows us to place the Universals in their corresponding group satisfactorily without being highly conditioned by the quantitative result obtained. Once there is greater availability of corpora, the result obtained for each Universal can be updated, and its position on the scale will change slightly. However, there should be no substantial changes within the group in which it is placed.

The GLU-M has made it possible to model the Universality of the claims analyzed in this doctoral thesis. Therefore, it has been possible to see its viability with concrete cases and to implement such a model, as exemplified in figure 23.1.

In this thesis, we have limited ourselves to proposing a model to be used to study the Universality of different rules in Linguistic Typology. However, we believe the present model can be adapted and applied to other questions of interest for Linguistic Typology. In particular, we believe that the trichotomous expression could be used to model the distribution of languages around a given linguistic type. In other words, we believe that the model could perfectly reflect, at the

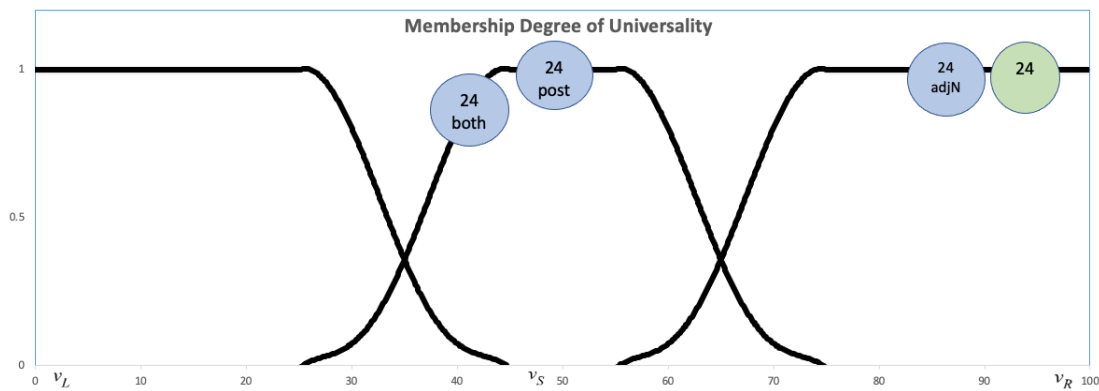


Figure 23.1: Example of a Trichotomous Expression for Language Universals

same time, the cluster of languages belonging to a given linguistic type and also be sufficiently gradual to contain the advantages shown above. This modeling of languages into types through the trichotomous expression could solve some of the existing limitations shared by the limitation of the formulation of Universals we have shown.

23.2 Formalization of Greenberg Universals

Another of the original contributions of the thesis is related to a bridge that we have drawn between Joseph Greenberg and Computational Linguistics. While it is true that there are some Universals analyzed computationally by other authors (although we do not agree on some aspects of their formalization as we have shown), we have not found a commitment to the holistic treatment of Greenberg's original proposal of 45 Universals from computation. Before our thesis, we did not know whether this revisiting of Greenberg from a different methodology would be feasible or appropriate. Therefore, we wanted to carry out this task.

23.2.1 Advantages and Limitations for a Computational Study of Greenberg Universals

Using Computational Linguistics tools to approach the 45 Universals formulated by Greenberg is motivated by some advantages compared to more traditional methodologies. Increasingly, we see a rise in interdisciplinarity and a clear interest in this new type of studies. The main aspects we can highlight are the following:

-
- **We can contrast data obtained by other methodologies with this quantitative computational data.** This kind of cross-checking helps to obtain much greater certainty about claims that are so complicated to certify. A multilevel approach, considering checks by means of UD, grammars, and tests, would be rich and ideal for dealing with the phenomena. If there is a coincidence in the results, we can reach a higher confidence, while results in different directions can invite us to study the phenomena in more depth.
 - **We can minimize the risk of bias in a second-hand source, with the consequences that this entails.** It is the data that sheds light on the phenomena. These data are testimonies of real texts that illustrate occurrences from different issuers, contexts, etc.
 - **The use of quantitative data allows for a very clear and traceable scientific methodology.** All the claims in this doctoral thesis are based on objective quantitative data that can be replicated following the described methodology. In addition, the data obtained are provided as an appendix to this thesis. These data, moreover, may be helpful for future researchers in other similar questions, which is one of the purposes of open science.
 - **The study of all the existing data in a text allows us to know borderline or non-prototypical structures, which traditionally have not been well represented.** In the traditional representation using grammars or other similar approaches, the linguistic information offered, logically, is based on a general description of languages. The Universals have also been proposed according to this traditional general knowledge, in which prepositional behavior is assimilated with an adposition preceding a noun, for example. In our study, we can study the relation of the adposition with other parts of speech and know the occurrences without this type of restriction, taking into account all the parts of speech at the same time. This will allow us to either extend the prototypical results to any other marginal situation or to know in more detail where the distribution is not fulfilled.
 - **A quantitative approach allows a much more fine-grained and richer weighting of the presence of the various rules.** The use of discrete categories may oversimplify a complex reality. Although in this thesis this idea has not had a central place (since it was a question of approaching Greenberg's claims and reviewing such categories), we have been able to point out some of the outputs that can be obtained through this typometrical approach. It is interesting to see, beyond whether a language is SOV or not, the behavior of a rigid SOV language (100% of occurrences) versus a SOV language with more flexibility (80% of occurrences, for example). We believe it is valuable to see if the correlations with other items

also follow this same distribution in which languages that are more rigid in type also show more rigid linguistic values. Ultimately, these checks allow us to look at the distributions of all the languages as a whole, seeing the shapes of the patterns created, which may reveal important and interesting restrictions at the linguistic level.

- **The study with this methodology also allows us to obtain a new kind of metrics.** Again, in this case, ideas like this appear in a residual form in the thesis and are not worked on since we have focused on verifying the Greenberg Universals (considering that it is the first step that should be done and has not been done). Therefore, it is also a future step we want to carry out. Thanks to the multiple computational data that can be obtained, it is possible to change the calculation of the direction of syntactic elements to a calculation of the distance between these, for example. Questions of this type can help to reveal very interesting behaviors of the world's languages.
- **We believe that easy, immediate, and free accessibility to the data of hundreds of languages is essential.** A study that requires the availability of hundreds of grammars (or the implementation of an in-depth study of social groups of more than 100 languages) requires resources that are often beyond the reach and decision-making power of the researcher. In addition, data management is fast, efficient, and orderly. We believe that the commitment to open science helps to reduce the significant gap (still existing) between researchers in different parts of the world, who may have very different access to information.

This series of advantages has invited us to try to represent the 45 Greenberg Universals computationally. The absence of studies like the one we have proposed makes it pioneering, and we must pay close attention to its feasibility. In other words, we will have to look at the limits of typological research with our proposed methodology.

In general, we must conclude that verifying the totality of Greenberg's 45 Universals is impossible. Multiple aspects condition this type of research, and that must be taken into account:

1. **Imbalance between language levels.** Greenberg Universals usually focus on the word order of syntactic elements. However, some also focus on aspects of morphosyntactic features or the morphological structure of words. There is also a Universal focus on intonation patterns.

The weight of syntax in computing is unquestionable. This makes it possible to consult Universals based on word order or morphosyntactic features. On the other hand, the Universal

relation to the intonation pattern and morphological structure of words cannot be checked. There is no such information annotated in UD. In other tools, it may be possible to have morphological information. However, these are not as well prepared as UD, so it is much more difficult to provide representative claims.

In summary, 34 of the 45 Universals in the 1963 publication can be formalized and treated. In other words, we have ascertained that it is not currently possible to conduct a total analysis of the aspects dealt with in Greenberg through UD.

2. Data annotation. We have detected some problems in annotating elements that affect three Universals. These problems cannot be solved directly by the researcher and must be solved from the tool, which is where these problems have been detected. We contribute to their detection and improvement by analyzing the causes detected. We consider that we can summarise it in two opposing errors:

- Firstly, we encounter problems related to missing data. These are usually originated when the Universal Dependencies annotation is mainly manual. The labelers or Treebank managers have chosen not to include a certain feature in their treebank. When that feature comes into play in some Universal, we can get misleading data. This lack of data can be constant or partial. Constant data missingness occurs when the annotators of a treebank do not use some of the tags that should be used in UD throughout the treebank, leaving it orphaned of that data. Partial absence refers to when the treebank does not include that label in a general way. However, some tagger forgets to apply it in some parts of the treebank, which generates examples of occurrences with this bias.
- Secondly, we encounter problems related to data overgeneration. These usually originated when the treebank was previously annotated with another schema and an automatic conversion to UD has been carried out. The parser assigns UD tags used in other situations to elements that should not have them, generating false occurrences of a particular element.

3. Specific structures. Another of the difficulties we have observed is when it comes to obtaining certain very specific and extensive occurrences. These can easily be proposed from the linguist's abstraction of the language he/she knows and studies. On the other hand, it is complicated to obtain spontaneous occurrences in real texts of concrete structures such as, for example, a noun with a numeral, an article determiner, and an adjective simultaneously (to see their order of appearance). In such cases, introspective testing is much more productive.

However, the absence of such structures can also be interesting to point out the complexity of these structures from a processing point of view, which conditions their occurrence.

4. **Language sampling.** The research we have carried out also allows an approach to the issue of the suitability (or not) of the languages available in this methodology we adopted. The conclusions we have reached are mixed. On the one hand, we must point out that the availability and richness of typological data in relation to different languages in Computational Linguistics is not comparable to the existing knowledge in grammars. Many languages still do not have labeled corpora and are, therefore, not represented. This makes the variety of sampling and the existing possibilities smaller in comparison.

On the other hand, however, we should be satisfied with the results obtained and obtainable. The 150 or so languages that have different corpora available for analysis show a decent approximation of the variety of the world's languages. We have enough information to offer a more substantial analysis than that proposed by Greenberg. Moreover, new languages are added to the repository year after year, allowing us to expand this knowledge. The growth is significant and constant.

5. **Corpora bias.** A classic limitation when studying languages through corpora is the influence that the type of corpus selected as representative of language behavior may have. This inherent limitation will always exist as a suspicion to a greater or lesser extent. This is also the case in the study we have proposed. However, we believe that precisely one of the advantages of the UD selection over other smaller repositories with comparable or parallel corpora is that there is an interesting variety of textual typologies. The variety of corpora and text types in most languages helps to obtain a broader and more representative picture. In addition, there is an effort to include texts that can be as unmarked as possible.

Some traditionally little-studied languages present only one corpus. Moreover, this corpus usually has a marked character, such as religious texts or popular narratives. In this case, we should point out that, on many occasions, this type of information is also what we obtain in other methodologies. That is to say, the information or perspective we currently have on some languages, even through grammar, is strongly conditioned by religious texts, for example, if this is the only accessible source available. In such a case, we cannot speak of a limitation of corpus typology exclusively in the methodology used, but this is relatively frequent in typological studies of under-represented languages.

In addition to showing the general benefits of a general computational approach to Greenberg Universals, we should also highlight as a contribution of this Ph.D. thesis the particular picture we have drawn for testing Greenberg Universals with particular tools and the implications that this entails.

We have not found in the literature any guide showing different tools for processing UD data, their instructions for use and their differences. Therefore, we have decided to carry out this research from scratch and offer it. Of the various tools analyzed, we considered Grew-Match to be the most appropriate for this purpose. We have also worked with Typometrics, if possible. In general, we wanted to show, step by step, how it is possible for a researcher in Linguistic Typology to carry out the type of search we have carried out to arrive at the final result we show. We have shown not only the Universals that cannot be verified from the UD data, but we have also offered a classification of the Universals from the different tools that can analyse them and those that cannot.

23.2.2 A Formalization of Greenberg's Universals

One of the outputs we contribute is the script with the formalization we make of the different linguistic elements contained in Greenberg's Universals. The originality lies in two different aspects.

1. On the one hand, in the literature, we have only found a few isolated formalizations referring to a specific aspect of one of Greenberg's Universals. Therefore, we offer a formalization for every Universal that can be formalized with the data available in UD. That is, we offer previously unpublished and unobserved formalizations of Greenberg Universals.
2. On the other hand, the formalizations we present in comparison with the existing ones are different. We have found a formal reading of the dominant basic word order concept that we discuss and argue against in all the proposals we have looked at. In addition to rejecting these proposals, we offer another type of formalization that better captures this phenomenon, essential in a lot of Greenberg's Universals.

To the formalizations of the structures mentioned by Greenberg, we must add the formalizations we have carried out of other structures that have allowed us to lay the groundwork for proposing other, different Universals. All the rules we have created and used to search for their occurrences in the different corpora have been made available in an open-access repository. Firstly, we facilitate and guarantee the falsifiability of the data we present. Secondly, since we believe in open science,

we provide the formalizations we have carried out in an accessible way in case they are of interest to scripts for future researchers in the field, who can make free use of them.

Greenberg's approach to Universals through formalizations has a fundamental advantage: the disambiguation of some problematic expressions. Using natural language (English) to capture any kind of structure may present some problems of correspondence or interpretation. There may not be an explicit reference to all the elements involved. In contrast, the formalizations we present are transparent with the elements they reference and are unambiguous.

23.3 Analysis of Greenberg Universals

We have offered a quantitative analysis based on various real texts in 143 languages. This has been carried out on two-thirds of the Universals formulated by Greenberg. The analysis is not intended to replace Greenberg and the methodological approach he promoted, which has lasted to the present day. However, we consider it valuable and relevant to promote different approaches to the same phenomenon in order to be able to capture the complex aspects involved in a multilevel way.

23.3.1 Testing Universals Computationally with Real Texts

We have carried out a review of Greenberg's Universals, which has allowed us to observe the behavior of the Universals in the broader set of languages. The texts of these different languages have served as the basis for observing the structures contained in Greenberg's Universals. We must highlight the great work of the linguist Joseph Greenberg, specifically about the Universals he has proposed. Such an achievement is not discovered in this doctoral thesis since, as we have seen, multiple linguists have used these claims for their research, and also the vast majority agree in pointing out that Greenberg is the father of modern Linguistic Typology. However, we believe this thesis can be considered further proof of the importance of formulated Universals. In this case, through a quantitative approach based on real texts, we have demonstrated the satisfactory behavior of most of these Universals. In other words, most Universals analyzed (21) are High-Universals. Eight of the Universals analyzed would be located in the Mid-Universal or Low-Universal groups and are not helpful information on the linguistic types they refer to. These Universals are usually related to the word order.

Not only do we classify the Universals and point out those that would not be appropriate according to the model and the data used, but we also provide a detailed explanation of the reasons for the poor performance of the Universal in the texts of the different languages. From what we have observed, the technical cause of the poor performance of some Universals can be attributed to two different reasons.

1. We can find Universals that correctly indicate a trend in the world's languages, but their formulation is not as precise as it should be. This means the data obtained, being strict with the formulation presented by Greenberg, does not allow a validity of the Universal. In many cases, we have been able to propose an alternative formulation (less extreme) that allows the conversion to High-Universal.
2. We find some Universals with multiple parts or possibilities. Some of these parts exhibit High-Universal values, but some other parts do not, which detracts from the overall results of the Universal. In these cases, we have believed that the most reasonable thing to do is to convert the Universal (or simplify it) so that it only points to the implication that we certify to be successful (High-Universal).

The data we have obtained allow us to confirm the claims of researchers who, by contrasting with other grammars, have pointed out the errors in Greenberg's formulations. In one case, we mark one of them as Mid-Universal, and the other directly formulate it in alternative terms given the impossibility of testing the claims as Greenberg proposed them. This alternative formulation of the Universal to be adequate is not usually observed in Greenberg's review studies of some Universals and what we do. Nor have we seen comprehensive revisions to the Universals such as the one we have done.

The methodology used is based on precise data to arrive at the final generality. For this reason, the mere verification of the result we have obtained with Greenberg includes a series of beneficial preliminary steps for Linguistic Typology. We obtain in detail the behavior of each language for the Universal. This type of information can be helpful in Linguistic Typology and general knowledge of one of the languages analyzed (if it has not been sufficiently studied).

We have also observed cases in which there is a disparity of criteria when assigning a linguistic type to some specific languages (more or less studied) in different grammars of the same language. This assignment, made by linguists, in some cases, admits the possible interpretation of the information. In our case, we rely on precise quantitative information based on occurrences in authentic

texts. This means that it is the data that guides us to resolve the doubts that may exist in a less data-based approach.

The information we provide individually about each language can be interesting for capturing traditionally invisible phenomena. It is necessary to remember that such data are circumscribed and conditioned by the available texts that have been processed. However, we believe that they adequately reflect the characteristics of the languages. Still today, the largest, most updated (or new) typological databases in the world, such as WALS or Grambank, present information based on the label of some selected grammar as a referent.

On the other hand, we have decided to create this information from UD. We can obtain a list of tags like the classical and prevailing databases, but we can also offer more detailed information about the phenomenon studied in each language. This information can be significant in understanding much less obvious implications, but it is conditioned by the presence (without being dominant) of some feature in such a language.

As an example, like other databases, we can offer that Spanish is a language whose dominant basic word order is SVO. However, we can also offer the percentage of structures we captured with the SOV, VSO, OVS, OSV, and VOS orders.

However, we also want to differentiate ourselves from current proposals in Computational Linguistics that leave aside the concept of linguistic type and categories and only show the quantitative distributions of languages. We are in favor of comparative work with Greenberg. This implies the need for comparable results. In order to do so, we have had to draw a bridge between the two methodologies to enable a conversion of the results into categories comparable to those of Greenberg.

In summary, this thesis checks Greenberg's Universals, further promoting their validity and validity. In turn, our contribution demonstrates that it is possible to carry out this type of analysis and, therefore, also contributes to validating studies of digitized linguistic corpora as a valuable tool for typological knowledge in the syntax of the world's languages.

23.3.2 Formulation of New Universals in Language

The doctoral thesis we present here has not focused exclusively on checking Greenberg's claims but has also aimed to investigate other types of linguistic correlations in parallel that may offer

significant results for Linguistic Typology. We contribute to Typology with the formulation of new Universals in Languages.

In total, we propose 20 new Universals. All of them have a clear High-Universal value. The starting point for formulating these Universals is Greenberg's claims. However, we have taken some actions that have modified them. Functionally, we can understand that there are two different procedures:

1. **Greenberg Universals with a result of Mid-Universal or Low-Universal.** Of the 8 Greenberg Universals that are not High-Universals, we converted 5 of them into High-Universals. We have been able to see in these cases that the Greenberg data and the selected languages have been able to mislead him, even though he correctly intuited some linkage between these linguistic elements. In many cases, we have detected that there is no correlation between the types proposed by Greenberg. However, there is the opposite version of the Universal, applicable to other types of languages.
2. **Greenberg Universals with a High-Universal result.** Most Greenberg Universals do not need an improvement in their formulation because they are already adequate. However, in the parallel checks we have carried out, we have been able to certify that there are other cases where the implication formulated by Greenberg also applies. We gain further linguistic knowledge by checking other cases where a linguistic type is closely related. Therefore, we consider that either Greenberg's Universal can be extended to more implications or that other implications can be formulated that show links between more specific linguistic elements not originally foreseen.

It is important to note what is involved in formulating these 20 concrete Universals.

- On the one hand, they should be highlighted as formulations that go beyond the mere verification of other rules already formulated by Greenberg. It is a more creative and novel contribution. We show unpublished associations as any other typologist could have done.
- On the other hand, they should be highlighted as formulations based on the idea of Joseph Greenberg. The Universals formulated arise in parallel to those formulated by Greenberg. They tend to extend Greenberg's claims, make them more precise, and so on. Any proposal made has a clear link with any of the Universals proposed by the American linguist.

The 20 Universals we provide show how our methodology is helpful in learning more about the behavior of the world's languages. Again, we do not understand Computational Linguistics as a substitute for grammars and other approaches. To a lesser extent at this point, when the resources and tools are not fully developed. However, we believe that, given the efficiency and the possibilities it opens up, the method used can be one of the first steps to start detecting trends that other levels of research can then confirm.

Analyzing different precise occurrences of specific structures has allowed us to verify a series of shared behaviors in languages that are very difficult to observe in much more general sources. This greater precision also allows for a more detailed knowledge of the rules shared between languages, which can facilitate further studies or applications.

UNIVERSITAT ROVIRA I VIRGILI

GLU - A GRADIENT MODEL FOR LANGUAGE UNIVERSALS. A COMPUTATIONAL REVISITATION OF GREENBERG UNIVERSALS

Antoni Brosa Rodríguez

24

Future Work

Contents

24.1	Future Research on Universality as a Gradient Concept	601
24.2	New Possibilities on the Formalization of Greenberg Universals	602
24.3	Future Work on the Analysis of Greenberg Universals	603

Once we have reviewed the main contributions of the doctoral thesis and their implications concerning the objectives set, we have seen that there are still many things to be done in the field. Some of these future works are avenues of research that have opened up to us during our research and which we have not delved into to not divert our interest from the main objective we had initially set ourselves. In other cases, future work is also linked to improving the above limitations. In summary, we can present different aspects related to future work in the different contributions of the thesis.

24.1 Future Research on Universality as a Gradient Concept

This thesis has presented the GLU-M as a possible model for studying Universals in Language. We have also carried out an application of this with Greenberg’s Universals. However, we believe that there are many other areas within Typology where it is possible to analyze the suitability of the proposed model:

- **Application to linguistic types.** We have proposed a model that deals with the degree of Universality of the rules in their relation to the world’s different languages. We believe

the model also allows for an adequate representation of linguistic types. Each of the languages analyzed exhibits a different degree of relevance to a feature. This gradation can be represented with our proposal, and the advantages and disadvantages of such modeling could be analyzed. In principle, the advantages we have pointed out in the characterization of Universals also apply to an analysis of this type.

- **Application to other linguistic fields.** We consider that the existence of the trichotomous expression should not be exclusively linked to Linguistic Typology. We can see how its proposal has an important place in the semantic modeling of quantifiers. Therefore, we are open to further work with the trichotomous expression in other linguistic aspects, as could be Greenberg quantifiers.

24.2 New Possibilities on the Formalization of Greenberg Universals

We have discussed the benefits and limitations of the Typology and Computation interdisciplinary approach. These, in turn, allow for further studies in the future, such as:

- **The study of Greenberg’s morphological Universals.** We have observed how it has proved impossible to formalize Greenberg’s purely morphological Universals with the UD annotation scheme. We believe that this lack of knowledge can be solved when progress in annotation and tagging in UD begins. We favor introducing annotations in the style of *interglossae*, as is usual in Linguistic Typology, to adequately capture the words’ structure. This kind of standardized information would allow a full knowledge of Greenberg’s Universals computationally without an answer to the Universal concerning intonational patterns.
- **The use of computational typological information for other disciplines.** We believe that the information we corroborated in this thesis or the new information we provided may be helpful for Computational Linguistics. However, we have not yet conducted any experiments, as this was not the main objective of the thesis. We have also detected new avenues of connection, not seen in the literature, between our work and other sub-disciplines, such as the study of linguistic complexity. We have understood the link between universality and complexity as an inversely proportional relationship. This has allowed us to use the information obtained to point out languages’ absolute and relative complexity. For absolute

complexity, we compare the results with the overall Universal result, showing whether the analyzed language is in the trend. For relative complexity, we can cross-reference the data of two given languages, for example, in order to be able to characterize the different linguistic structures shared or not.

- **The proposal of more specific formalizations.** The proposed analysis allows the efficient testing of concrete structures. About Universal 1, for example, we could check all the word order possibilities within a relative clause (and to see if there are any limitations by the linguistic type). Alternatively, for example, we could filter the results of pronouns by isolating the ones relative to the rest to see how the different pronoun types can behave in languages and to check for any invariability. It is a complex but interesting study we would wish to develop in the next months or years.

24.3 Future Work on the Analysis of Greenberg Universals

The analysis we have carried out in our research offers new lines of future research and opportunities to return to this topic later. Among other aspects, we can talk about the following:

- **Contrast of the results obtained in the analysis.** We have indicated that the sampling allowed by UD version 2.11 is not as rich and varied as desirable. Although we have considered such sampling sufficient, returning to the elements analyzed in future versions would be interesting. In versions with more languages and a greater variety, we will still be able to establish a comparison with our data and see how valuable and meaningful the data we have worked with are.

Thanks to the model used, the weight of this sampling is minimized, and the changes would not be significant. However, this work should be done in the future to establish it in a rigorous and univocal way.

- **Study of Greenberg's Universals based on new metrics.** Joseph Greenberg focused on the direction of syntactic elements as the analyzable aspect to form meaningful linguistic types. We have shown how the quantitative study of corpora allows linguistic knowledge to be gained through much more novel metrics such as Dependency Length.

The main objective of this thesis was to test the dependency direction to establish a clear comparison with Greenberg's claims. The fact that we want to keep to this first step of the

thesis in order not to disperse its content does not mean that we do not consider that, in the future, we can incorporate new metrics. In our opinion, these allow us, fundamentally, to go further in the final existing knowledge in Linguistic Typology: the explanation of the latent functional criteria in the composition of the world's languages.

- **Development of new Universals.** One of the clear contributions of this doctoral thesis is the formulation of new Universals based on the tools and methodology we have shown. We have managed to propose 20 new Universals linked to Greenberg's proposals. In the case of two Universals, we have observed a possible extension of these Universals or conversion into an implicational hierarchy (Universals 36 and 43). However, we have not been able to go into this idea in-depth due to the lack of space to address it rigorously. Checks must be carried out using other methodologies to help confirm this suspicion and diachronic data should also support the claim. If we could confirm these two implicational hierarchies, the proposed methodology would be confirmed as a good option for prediction or first-level detection of linguistic correlations.
- **Typometrics.** Our proposed analysis of Greenberg's Universals has also included, where possible, graphs produced with the Typometrics tool. Although we have provided an introduction to the tool and the interpretation of the results in these terms, we have not done so in an in-depth and detailed manner. We considered that this was not the main objective of the thesis and could divert attention from the essential purposes. However, we have considered Typometrics as a useful tool that deserves further development in the future to provide typological characterizations as we have exemplified with some Universals.
- **Research on other Universals using the same methodology.** We believe that the methodology we have shown and demonstrated allows for the future development of computational research on other Universals. These may even be at another level, such as the lexicon.
- **Application of the results in second language learning studies.** The typological information obtained can be of great interest when organizing the curriculum of a foreign language classroom. Typological information on the source language and the target language can help better organize the presentation and teaching content. We believe that there is still some way to go in the Universals.

Appendix: Script

We provide the necessary data as an appendix to facilitate the replicability of the study we have carried out in this thesis. We believe in open science and have chosen to distribute it in an open-access repository. The provided script shows all the formalizations corresponding to the Greenberg universals and other formalizations related to the respective universals and more detailed information about them. In order to reach the conclusions shown in this thesis, one has to apply the query in the script to the first 1000 treebanks of all Universal Dependencies corpora (2.11). However, in the body of the thesis, we have already given the essential information to replicate such data, with a sample of the formalizations concerned and their explanation.

Open Access Repository DOI Link: <https://doi.org/10.34810/data986>

UNIVERSITAT ROVIRA I VIRGILI

GLU - A GRADIENT MODEL FOR LANGUAGE UNIVERSALS. A COMPUTATIONAL REVISITATION OF GREENBERG UNIVERSALS

Antoni Brosa Rodríguez

Bibliography

- Abraham, W. (1989). Language Universals: The Chomskyan Approach vs. Greenberg's Typological Approach. *Belgian Journal of Linguistics*, 4, 9-25. doi: 10.1075/bjl.4.02abr
- Allan, K. (1987). Hierarchies and the coince of left conjuncts (with particular attention to English). *Journal of Linguistics*, 23(1), 51-77.
- Allen, J., & Core, M. (1997). *Draft of DMSL: Dialogue Act Markup in Several Layers*. Retrieved from <http://www.cs.rochester.edu/research/cisd/resources/damsl/RevisedManual/RevisedManual.html>
- Alzetta, C., Dell'Orletta, F., Montemagni, S., & Venturi, G. (2019). Inferring quantitative typological trends from multilingual treebanks. A case study. *Lingue e Linguaggio*, 18(2), 209-242. doi: 10.1418/95391
- Anderson, S. R. (2005). *How many languages are there in the world?* Retrieved from http://www.lsadc.org/pdf_files/howmany.pdf
- Arcodia, G. F. (2022a). La tipologia morfologica. In N. Grandi & C. Mauri (Eds.), *La tipologia linguistica: unità e diversità nelle lingue del mondo* (p. 178-198). Roma: Carocci Editore.
- Arcodia, G. F. (2022b). La tipologia sintattica. In N. Grandi & C. Mauri (Eds.), *La tipologia linguistica: unità e diversità nelle lingue del mondo* (p. 199-218). Roma: Carocci Editore.
- Asher, R. E. (1985). *Tamil: Descriptive Grammars*. London: Croom Helm.
- Austin, J. L. (1962). *How to do things with words*. Oxford: Oxford University Press.
- Bach, E., & Chao, W. (2009). On Semantic Universals and Typology. In M. Christiansen, C. Collins, & S. Edelman (Eds.), *Language Universals* (p. 152-173). New York: Oxford University Press.
- Bakker, D. (2008). INFER: Inferring implications from the WALS database. *Sprachtypologie un Universalienforschung*, 61(3), 186-198.
- Bakker, D. (2010). Language Sampling. In J. J. Song (Ed.), *The Oxford Handbook of Linguistic Typology* (p. 1-26). Oxford: Oxford University Press. doi: 10.1093/oxfordhb/9780199281251.013.0007
- Barth, D., & Schnell, S. (2022). *Understanding Corpus Linguistics*. New York: Routledge.
- Becker, L. (2021). *Articles in the World's Languages*. Berlin: De Gruyter.

- Bell, A. (1978). Language samples. In J. Greenberg, C. Ferguson, & E. Moravcsik (Eds.), *Universals of Human Language* (p. 123-156). Stanford: Stanford University Press.
- Belyaev, O. (2008). Stability of language features: A comparison of WALS and JM typological databases. In *Cognitive Modeling in Linguistics*. Becici, Montenegro.
- Bentz, C., Gutierrez-Vasques, X., Sozinova, O., & Samardžić, T. (2022). Complexity trade-offs and equi-complexity in natural languages: a meta-analysis. *Linguistics Vanguard*, 8, 1-17. doi: 10.1515/lingvan-2021-0054
- Berlin, B., & Kay, P. (1969). *Universality and Evolution of Basic Color Terms*. Berkeley: University of California Press.
- Bever, T. G. (2009). Minimalist Behaviorism: The Role of the Individual in Explaining Language Unviersals. In M. Christiansen, C. Collins, & S. Edelman (Eds.), *Language Universals* (p. 99-125). New York: Oxford University Press.
- Biberauer, T. (2012). Verb second in Afrikaans: Is this a unitary phenomenon? *Stellenbosch Papers in Linguistics*, 34(0). doi: 10.5774/34-0-34
- Bickel, B. (2007). Typology in the 21st century: Major current developments. *Linguistic Typology*, 11(1), 239-251. doi: 10.1515/LINGTY.2007.018
- Bickel, B. (2010). Absolute and Statistical Universals. In P. Colm Hogan (Ed.), *The Cambridge Encyclopedia of the Language Sciences* (p. 77-79). Cambridge: Cambridge University Press.
- Bickel, B. (2015). Distributional Typology: Statistical Inquires into the Dynamics of Linguistic Diversity. In *The Oxford Handbook of Linguistic Analysis* (p. 1-32). Oxford University Press. doi: 10.1093/oxfordhb/9780199677078.013.0046
- Bickel, B. (2017a). Areas and universals. In R. Hickey (Ed.), *The Cambridge Handbook of Areal Linguistics* (p. 40-55). Cambridge: Cambridge University Press. doi: 10.1017/9781107279872.004
- Bickel, B. (2017b). Linguistic diversity and universals. In N. Enfield, P. Kockelman, & J. Sidnell (Eds.), *Cambridge Handbook of Linguistic Anthropology* (Vol. 16, p. 102-127). Cambridge: Cambridge University Press. doi: 10.1418/88241
- Bickel, B., & Nichols, J. (2003). *Typological Enclaves*. Retrieved from [http://www.unileipzig.de/\\$\sim\\$autotyp/download/index.html](http://www.unileipzig.de/$\sim$autotyp/download/index.html)
- Bickel, B., & Nichols, J. (2013). *The Autotyp genealogy and geography database 2013 release*. Retrieved 2022-10-22, from <http://www.autotyp.uzh.ch/available.html>
- Bisang, W. (2010). Word Classes. In J. J. Song (Ed.), *The Oxford Handbook of Linguistic Typology* (p. 1-18). Oxford: Oxford University Press.
- Blasi, D. E., Wichmann, S., Hammarström, H., Stadler, P. F., & Christiansen, M. (2016). Sound-

- meaning association biases evidenced across thousands of languages. *Proceedings of the National Academy of Sciences of the United States of America*, 113, 10818-10823. doi: 10.1073/pnas.1605782113
- Bloomfield, L. (1933). *Language*. London: George Allen & Unwin Ltd.
- Blum-Kulka, S. (1981). The Study of Translation in View of New Developments in Discourse Analysis: The Problem of Indirect Speech Acts. *Poetics Today*, 2(4), 89-95. doi: 10.2307/1772488
- Blum-Kulka, S., & Olshtain, E. (1984). Requests and apologies: A cross-cultural study of speech act realization patterns (CCSARP). *Applied Linguistics*, 5(3), 196-213. doi: 10.1093/applin/5.3.196
- Bobaljik, J. D. (2012). *Universals in Comparative Morphology: Suppletion, Superlatives, and the Structure of Words*. Massachusetts: MIT Press.
- Boeckx, C., Fodor, J., & Glerman, L. (2009). Round Table: Language Universals: Yesterday, Today, and Tomorrow. In M. Piattelli-Palmarini, J. Uriagereka, & P. Salaburu (Eds.), *Of minds and language: A dialogue with Noam Chomsky in the Basque Country* (p. 195-220). Oxford: Oxford University Press.
- Booij, G. (2009). Lexical Integrity as a Formal Universal: A Constructionist View. In S. Scalise, E. Magni, & A. Bisetto (Eds.), *Universals of Language Today* (p. 83-100). Bologna: Springer.
- Bowerman, M. (2010). Linguistic Typology and First Language Acquisition. In J. J. Song (Ed.), *The Oxford Handbook of Linguistic Typology* (p. 1-17). Oxford: Oxford University Press.
- Brosa-Rodríguez, A., & Jiménez-López, M. D. (2023). A Typometrical Study of Greenberg's Linguistic Universal 1. In R. Mehmood et al. (Eds.), *Lecture Notes in Networks and Systems: Distributed Computing and Artificial Intelligence* (p. 186-196). Springer. doi: 10.1007/978-3-031-38318-2_19
- Brosa-Rodríguez, A., & Jiménez-López, M. D. (2024). Quantifying basic colors' salience from cross-linguistic corpora. *Color Research and Application*(49), 34-50. doi: 10.1002/col.22899
- Brown, D. (2010). Morphological Typology. In J. J. Song (Ed.), *The Oxford Handbook of Linguistic Typology* (p. 1-14). Oxford: Oxford University Press.
- Brown, P., & Levinson, S. (1988). *Politeness: Some Universals in Language Usage*. Cambridge: Cambridge University Press. doi: 10.2307/3587263
- Burquest, D. A. (1973). *A grammar of Angas* (Unpublished doctoral dissertation). University of California.
- Burton, S., & Matthewson, L. (2015). Targeted Construction Storyboards in Semantic Fieldwork. In R. Bochnak & L. Matthewson (Eds.), *Semantic Fieldwork Methodology* (p. 135-156).

- Oxford: Oxford University Press.
- Bybee, J. (2009). Language Universals and Usage-Based Theory. In M. Christiansen, C. Collins, & S. Edelman (Eds.), *Language Universals* (p. 17-39). New York: Oxford University Press.
- Bybee, J. (2010). Markedness: Iconicity, Economy, and Frequency. In J. J. Song (Ed.), *The Oxford Handbook of Linguistic Typology* (p. 1-11). Oxford: Oxford University Press.
- Campbell, L., Albany, S., Bubenik, V., & Saxon, L. (1988). Word Order Universals: Refinements and Clarifications. *Canadian Journal of Linguistics*, 33(3), 209-230. doi: 10.1017/s0008413100012962
- Carmichael, C. (2010). Universals. *Philosophical Studies*, 150(3), 373-389. doi: 10.1007/s11098-009-9417-3
- Casagrande, J. B. (1963). Language Universals in Antropological Perspective. In J. Greenberg (Ed.), *Universals of Language* (p. 220-235). Cambridge, Massachussetts: The M.I.T. Press.
- Cela-Conde, C. J., & Marty, G. (1997). Noam Chomsky's Minimalist Program and the Philosophy of Mind. An Interview. *Syntax*, 1(1), 19-36. doi: 10.1111/1467-9612.00002
- Chen, X., & Gerdes, K. (2017). Classifying Languages by Dependency Structure Typologies of Delexicalized Universal Dependency Treebanks. In S. Montemagni & J. Nivre (Eds.), *Proceedings of the Fourth International Conference on Dependency Linguistics* (p. 54-63).
- Chen, X., Gerdes, K., Kahane, S., & Courtin, M. (2021). The Co-Effect of Menzerath-Altmann Law and Heavy Constituent Shift in Natural Languages. In *Proceedings of Qualico 2021* (p. 1-15).
- Choi, H. S. (2021). *Vérification d'universaux linguistiques sur des corpus multilingues annotés en syntaxe de dépendances* (Unpublished doctoral dissertation). Sorbonne Université, Paris.
- Choi, H. S., Guillaume, B., & Fort, K. (2021). Corpus-based Language Universals Analysis using Universal Dependencies. In R. Cech & X. Chen (Eds.), *Proceedings of the Second Workshop on Quantitative Syntax* (p. 1-15).
- Choi, H. S., Guillaume, B., Fort, K., & Perrier, G. (2021). Investigating Dominant Word Order on Universal Dependencies with Graph Rewriting. In *International Conference Recent Advances in Natural Language Processing, RANLP* (p. 281-290). Incoma Ltd. doi: 10.26615/978-954-452-072-4_033
- Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge: The M.I.T. Press. doi: 10.1007/978-94-017-2239-1_2
- Christiansen, M., & Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences*, 31(5), 489-508.
- Christiansen, M., & Chater, N. (2009). The myth of language universals and the myth of universal

- grammar. *Behavioral and Brain Sciences*, 32(5), 452-453. doi: 10.1017/S0140525X09990641
- Christiansen, M., Collins, C., & Edelman, S. (2009). *Language Universals*. New York: Oxford University Press.
- Cinque, G. (2005). Deriving Greenberg's Universal 20 and its exceptions. *Linguistic Inquiry*, 36(3), 315-332. doi: 10.1162/0024389054396917
- Cinque, G. (2009). The Fundamental Left-Right Asymmetry of Natural Languages. In S. Scalise, E. Magni, & A. Bisetto (Eds.), *Universals of Language Today* (p. 165-184). Bologna: Springer.
- Cinque, G. (2013a). Greenberg's Universal 23 and SVO languages. In G. Cinque (Ed.), *Typological Studies* (p. 99-103). New York: Routledge.
- Cinque, G. (2013b). *Typological Studies: Word Order and Relative Clauses*. New York / London: Routledge.
- Clark, A., & Misyak, J. B. (2009). Language, Innateness, and Universals. In M. Christiansen, C. Collins, & S. Edelman (Eds.), *Language Universals* (p. 253-260). New York: Oxford University Press.
- Clark, H., & Bangerter, A. (2004). Changing Ideas about Reference. In I. A. Noveck & D. Sperber (Eds.), *Experimental Pragmatics* (p. 25-49). Houndmills: Palgrave Macmillan.
- Coke, R., King, B., & Radev, D. R. (2016). Classifying Syntactic Regularities for Hundreds of Languages. In *Computer Research Repository* (p. 1-10).
- Cole, P. (1982). *Imbabura Quechua: Descriptive Grammars*. Amsterdam: North Holland.
- Coloma, G. (2017). Complexity trade-offs in the 100-language WALS sample. *Language Sciences*, 59, 148-158.
- Comrie, B. (1989). *Language Universals and Linguistic Typology*. Chicago: The University of Chicago Press.
- Comrie, B. (2006). Syntactic typology. In R. Mairal & J. Gil (Eds.), *Linguistic Universals* (p. 130-154). Cambridge: Cambridge University Press.
- Comrie, B. (2020). Revisiting Greenberg's "Generalizations about Numeral Systems" (1978). *Journal of Universal Language*, 21(2), 43-84. doi: 10.22425/jul.2020.21.2.43
- Comrie, B., & Keenan, E. L. (1979). Noun Phrase Accessibility Revisted. *Language*, 55(3), 649-664.
- Comrie, B., & Smith, N. (1977). *The Lingua Descriptive Studies Questionnaire*. Retrieved from https://www.eva.mpg.de/lingua/tools-at-lingboard/questionnaire/lingua-descriptive_description.php
- Cook, R. T. (2002). Vagueness and Mathematical Precision. *Mind*, 111(442), 225-247.

- Coopmans, P. (1983). Language universals and linguistic typology. *Journal of Linguistics*, 19(2), 455-473. doi: 10.1017/S0022226700007817
- Coopmans, P. (1984). Surface Word-Order Typology and Universal Grammar. *Language*, 60(1), 55-69.
- Cooreman, A., & Goyvaerts, D. (1980). Universals in Human Language. A Historical Perspective. *Revue belge de philologie et d'histoire*, 58(3), 615-638. doi: 10.3406/rbph.1980.3295
- Corbett, G. G. (2009). Universals and Features. In S. Scalise, E. Magni, & A. Bisetto (Eds.), *Universals of Language Today* (p. 129-144). Bologna: Springer.
- Core, M., & Schubert, L. K. (1996). *Dialog Parsing in the TRAINS System*. Rochester: University of Rochester.
- Cowgill, W. (1963). A Search for Universals in Indo-European Diachronic Morphology. In J. Greenberg (Ed.), *Universals of Language* (p. 91-113). Cambridge, Massachusetts: The M.I.T. Press.
- Cristofaro, S. (2010). Language Universals and Linguistic Knowledge. In J. J. Song (Ed.), *The Oxford Handbook of Linguistic Typology*. Oxford: Oxford University Press. doi: 10.1093/oxfordhb/9780199281251.013.0013
- Cristofaro, S. (2019). From synchronically-oriented typology to source-oriented typology: Typological universals in diachronic perspective. In C. Gianollo & C. Mauri (Eds.), *CLUB Working Papers on Linguistics* (p. 52-68). Bologna: Alma Mater Studiorum Università di Bologna.
- Cristofaro, S. (2022). Spiegazioni sincroniche e diacroniche per gli universali tipologici. In N. Grandi & C. Mauri (Eds.), *La tipologia linguistica: unità e diversità nelle lingue del mondo* (p. 97-120). Roma: Carocci Editore.
- Croft, W. (2003). *Typology and Universals*. Cambridge: Cambridge University Press. doi: 10.2307/416749
- Croft, W. (2009a). Methods for Finding Language Universals in Syntax. In S. Scalise, E. Magni, & A. Bisetto (Eds.), *Universals of Language Today* (p. 145-164). Bologna: Springer.
- Croft, W. (2009b). Syntax is more diverse, and evolutionary linguistics is already here. *Behavioral and Brain Sciences*, 32(5), 453-454.
- Croft, W., Nordquist, D., Looney, K., & Regan, M. (2017). Linguistic Typology Meets Universal Dependencies. In M. Dickinson, J. Hajic, S. Kubler, & A. Przepiorkowski (Eds.), *Proceedings of the conference on Treebanks and Linguistic Theories* (p. 63-75).
- Croft, W., & Poole, K. (2008b). Multidimensional scaling and other techniques for uncovering universals. *Theoretical Linguistics*, 34(1), 75-84. doi: 10.1515/THLI.2008.007
- Croft, W., & Poole, K. T. (2008a). Inferring universals from grammatical variation: Mul-

- tidimensional scaling for typological analysis. *Theoretical Linguistics*, 34(1), 1-37. doi: 10.1515/THLI.2008.001
- Crystal, D. (2008). *A dictionary of linguistics and phonetics*. Oxford: Blackwell Publishing.
- Cysouw, M. (2003). Against implicational universals. *Linguistic Typology*, 7(1), 89-101.
- Cysouw, M. (2009). Using the World Atlas of Language Structures. *Language Typology and Universals*, 61(3), 1-6.
- Cysouw, M., & Wohlgemuth, J. (2010). The other end of universals: theory and typology of rare. In J. Wohlgemuth & M. Cysouw (Eds.), *Rethinking Universals: How Rarities affect Linguistic Theory* (p. 1-10). Berlin: De Gruyter Mouton.
- Dabrowska, E. (2015). What exactly is Universal Grammar, and has anyone seen it? *Frontiers in Psychology*, 6, 1-17. doi: 10.3389/fpsyg.2015.00852
- Dahl, Ö. (1980). Is linguistics empirical? A critique of Esa Itkonen's Linguistics and Metascience. In T. Perry (Ed.), *Evidence and Argumentation in Linguistics* (p. 133-145). Berlin/New York: Mouton de Gruyter. doi: 10.1515/9783110848854-008
- Daniel, M. (2010). Linguistic Typology and the Study of Language. In J. J. Song (Ed.), *The Oxford Handbook of Linguistic Typology* (p. 1-16). Oxford: Oxford University Press.
- Daumé III, H., & Campbell, L. (2007). A Bayesian model for discovering typological implications. In A. Zaenen & A. van den Bosch (Eds.), *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics* (p. 65-72).
- Davies, J. (1981). *Kobon: Descriptive Grammars*. Amsterdam: North Holland.
- De Marneffe, M.-C., Connor, M., Silveira, N., Bowman, S. R., Dozat, T., & Manning, C. D. (2013). More constructions, more genres: Extending Stanford Dependencies. In E. Hajicova, K. Gerdes, & L. Wanner (Eds.), *Proceedings of the Second International Conference on Dependency Linguistics* (p. 187-196).
- De Marneffe, M.-C., & Manning, C. D. (2008). The Stanford typed dependencies representation. In J. Bos et al. (Eds.), *COLING Workshop on Cross-framework and Cross-domain Parser Evaluation*.
- De Vries, M. (2005). The Fall and Rise of Universals on Relativization. *Journal of Universal Language*, 6, 125-157.
- De Vries, M. (2018). Relative Clauses in Syntax. In *Oxford Research Encyclopedia of Linguistics* (p. 1-30). Oxford: Oxford University Press.
- Delfitto, D. (2009). Universals and Semantics. In S. Scalise, E. Magni, & A. Bisetto (Eds.), *Universals of Language Today* (p. 209-224). Bologna: Springer.
- De Marneffe, M.-C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., & Manning,

- C. D. (2014). Universal Stanford Dependencies: A cross-linguistic typology. In N. Calzolari et al. (Eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*.
- De Marneffe, M. C., Manning, C. D., Nivre, J., & Zeman, D. (2021). Universal dependencies. *Computational Linguistics*, 47(2), 255-308. doi: 10.1162/COLI_a_00402
- Diessel, H., & Hetterle, K. (2011). Causal clauses: a cross-linguistic investigation of their structure, meaning, and use. In P. Siemund (Ed.), *Linguistic Universals and Language Variation* (p. 23-54). Berlin: De Gruyter Mouton.
- Droganova, K., & Zeman, D. (2019). Towards Deep Universal Dependencies. In M. Candito, K. Evang, S. Oepen, & D. Seddah (Eds.), *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)* (p. 1-9).
- Dryer, M. (1987). *Final Report to Social Sciences and Humanities Council of Canada on Research Grant No. 410-85-0540: A statistical study of word order universals*. Canada: Council of Canada.
- Dryer, M. (1988a). Object-Verb Order and Adjective-Noun Order: Dispelling a Myth. *Lingua*, 74, 185-217.
- Dryer, M. (1988b). Universals of negative position. In Hammond, Moravcsik, & Wirth (Eds.), *Studies in syntactic typology* (p. 93-124). Amsterdam: John Benjamins.
- Dryer, M. (1989). Large linguistic areas and language sampling. *Studies in Language*, 13(2), 257-292. doi: 10.1075/sl.13.2.03dry
- Dryer, M. (1992). The Greenbergian Word Order Correlations. *Language*, 68(1), 81-138.
- Dryer, M. (1995). Frequency and pragmatically unmarked word order. In P. Downing & M. Noonan (Eds.), *Word Order in Discourse* (p. 105-135). Amsterdam/Philadelphia: John Benjamin Publishing. doi: 10.1075/tsl.30.06dry
- Dryer, M. (1997). On the Six-Way Word Order Typology. *Studies in Language*, 21(1), 69-103. doi: 10.1075/sl.21.1.04dry
- Dryer, M. (1998). Why Statistical Universals are Better than Absolute Universals. In *Papers from the 33rd Annual Meeting of the Chicago Linguistics Society* (p. 1-23). Chicago: Chicago Linguistics Society.
- Dryer, M. (2005). *Genealogical language list*. Retrieved from <http://blog.wals.info/errata-in-the-printed-edition-of-2005/>
- Dryer, M. (2006). *The order of demonstrative, numeral, adjective and noun: an alternative to Cinque*. Retrieved 2022-10-20, from http://exadmin.matita.net/uploads/pagine/1898313034_cinqueH09.pdf

- Dryer, M. (2009). The Branching Direction Theory of Word Order Correlations Revisited. In S. Scalise, E. Magni, & A. Bisetto (Eds.), *Universals of Language Today* (p. 185-208). Bologna: Springer.
- Dryer, M. (2011). The evidence for word order correlations. *Linguistic Typology*, 15(2), 335-380. doi: 10.1515/LITY.2011.024
- Dryer, M. (2013). On the Six-Way Word Order Typology, Again. *Studies in Language*, 37(2), 267-301. doi: 10.1075/sl.37.2.02dry
- Dryer, M. (2019). On the order of Demonstrative, Numeral, Adjective, and Noun. *Language*, 94(4), 798-833.
- Dryer, M., & Haspelmath, M. (2013). *The World Atlas of Language Structures Online*. Retrieved from <http://wals.info>
- Dunn, M., Greenhill, S., Levinson, S., & Gray, R. (2011). Evolved structure of language shows lineage-specific trends in word-order universals. *Nature*, 473(7345), 79-82. doi: 10.1038/nature09923
- Eberhard, D., Simons, G., & Fennig, C. (2023). *Ethnologue: Languages of the World*. Retrieved 2023-08-22, from <http://www.ethnologue.com>
- Eckman, F. (2010). Linguistic Typology and Second Language Acquisition. In J. J. Song (Ed.), *The Oxford Handbook of Linguistic Typology* (p. 1-11). Oxford: Oxford University Press.
- Eckman, F., Bell, L., & Nelson, D. (1988). On the Generalization of Relative Clause Instruction in the Acquisition of English as a Second Language. *Applied Linguistics*, 9, 1-20.
- Égré, P., & Klinedinst, N. (2011). *Vagueness and Language Use*. London: Palgrave Macmillan. doi: 10.1057/9780230299313
- Epps, P. (2010). Linguistic Typology and Language Documentation. In J. J. Song (Ed.), *The Oxford Handbook of Linguistic Typology* (p. 1-10). Oxford: Oxford University Press.
- Escandell-Vidal, V. (1996). Towards a cognitive approach to politeness. *Language Sciences*, 18(3-4), 629-650. doi: 10.1016/S0388-0001(96)00039-3
- European Science Foundation. (1993). *Eurotyp Questionnaires*. Retrieved 2022-09-19, from https://www.eva.mpg.de/lingua/tools-at-lingboard/questionnaire/eurotyp_description.php
- Evans, N. (2010). Semantic Typology. In J. J. Song (Ed.), *The Oxford Handbook of Linguistic Typology* (p. 1-20). Oxford: Oxford University Press.
- Evans, N., & Levinson, S. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, 32(5), 429-492. doi: 10.1017/S0140525X0999094X

-
- Ferguson, C. (1963). Assumptions about Nasals: A Sample Study in Phonological Universals. In J. Greenberg (Ed.), *Universals of Language* (p. 42-47). Cambridge, Massachusetts: The M.I.T. Press.
- Ferguson, C. (1978a). Historical Background of Universals Research. In J. Greenberg, C. Ferguson, & E. Moravcsik (Eds.), *Universals of Human Language* (p. 7-31). Stanford: Stanford University Press.
- Ferguson, C. (1978b). Talking to Children: A Search for Universals. In J. Greenberg, C. Ferguson, & E. Moravcsik (Eds.), *Universals of Human Language* (p. 203-224). Stanford: Stanford University Press.
- Ferrer i Cancho, R. (2003). *Language: universals, principles and origins* (Unpublished doctoral dissertation). Universitat Politècnica de Catalunya.
- Fox, B. A. (1987). The noun phrase accessibility hierarchy reinterpreted: subject primacy or the absolute hypothesis? *Language*, 63(4), 856-870.
- Fukada, A., & Asato, N. (2004). Universal politeness theory: Application to the use of Japanese honorifics. *Journal of Pragmatics*, 36(11), 1991-2002. doi: 10.1016/j.pragma.2003.11.006
- Futrell, R., Levy, R. P., & Gibson, E. (2020). Dependency locality as an explanatory principle for word order. *Language*, 96(2), 371-412. doi: 10.1353/lan.2020.0024
- Futrell, R., Mahowald, K., & Gibson, E. (2015). Quantifying Word Order Freedom in Dependency Corpora. In J. Nivre & E. Hajicova (Eds.), *Proceedings of the Third International Conference on Dependency Linguistics* (p. 91-100).
- Gajovic, V., Kerkez, M., & Kocovic, J. (2018). Modeling and simulation of logistic processes: risk assessment with a fuzzy logic technique. *Simulation: Transactions of the Society for Modeling and Simulation International*, 94(6), 507-518.
- Galluzzo, G., & Loux, M. J. (2018). *The Problem of Universals in Contemporary Philosophy*. Cambridge: Cambridge University Press.
- Garrette, D. (2015). *Inducing Grammars from Linguistic Universals and Realistic Amounts of Supervision* (Unpublished doctoral dissertation). The University of Texas at Austin.
- Georgi, R., Xia, F., & Lewis, W. (2010). Comparing language similarity across genetic and typologically-based groupings. In C.-R. Huang & D. Jurafsky (Eds.), *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)* (p. 385-393).
- Gerdes, K., Guillaume, B., Kahane, S., & Perrier, G. (2018). SUD or Surface-Syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD. In *Proceedings of the Second Workshop on Universal Dependencies* (p. 66-74). Association for Computational Linguistics.
- Gerdes, K., Guillaume, B., Kahane, S., & Perrier, G. (2019). Improving Surface-syntactic Universal

- Dependencies (SUD): surface-syntactic relations and deep syntactic features. In M. Candito, K. Evang, S. Oepen, & D. Seddah (Eds.), *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)* (p. 1-7).
- Gerdes, K., Guillaume, B., Kahane, S., & Perrier, G. (2022). Starting a new treebank? Go SUD! Theoretical and practical benefits of the Surface-Syntactic distributional approach. In N. Mazziotta & S. Mille (Eds.), *SyntaxFest Depling 2021 - 6th International Conference on Dependency Linguistics* (p. 35-46).
- Gerdes, K., & Kahane, S. (2016). Dependency Annotation Choices: Assessing Theoretical and Practical Issues of Universal Dependencies. In A. Friedrich & K. Tomanek (Eds.), *Proceedings of LAW X - The 10th Linguistic Annotation Workshop* (p. 131-140).
- Gerdes, K., Kahane, S., & Chen, X. (2019). Rediscovering Greenberg's Word Order Universals in UD. In M. Candito, K. Evang, S. Oepen, & D. Seddah (Eds.), *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)* (p. 124-131). doi: 10.18653/v1/W19-8015
- Gerdes, K., Kahane, S., & Chen, X. (2021). Typometrics: From implicational to quantitative universals in word order typology. *Glossa*, 6(1). doi: 10.5334/gjgl.764
- Gil, D. (1997). The Structure of Riau Indonesian. *Nordic Journal of Linguistics*, 17, 179-200.
- Gilquin, G., & Gries, S. T. (2009). Corpora and experimental methods: A state-of-the-art review. *Corpus Linguistics and Linguistic Theory*, 5(1), 1-26. doi: 10.1515/CLLT.2009.001
- Graffi, G. (2010). The Pioneers of Linguistic Typology: From Gabelentz to Greenberg. In J. J. Song (Ed.), *The Oxford Handbook of Linguistic Typology* (p. 1-11). Oxford: Oxford University Press.
- Grandi, N. (2020). *Fondamenti di tipologia linguistica*. Roma: Carocci.
- Grandi, N., & Mauri, C. (2022). *La tipologia linguistica: unità e diversità nelle lingue del mondo*. Roma: Carocci Editore.
- Greenberg, J. (1960). A Quantitative Approach to the Morphological Typology of Language. *International Journal of American Linguistics*, 26(3), 178-194.
- Greenberg, J. (1963). *Universals of Language*. Cambridge, Massachusetts: The M.I.T. Press.
- Greenberg, J. (1978a). Diachrony, Synchrony and Language Universals. In J. Greenberg, C. Ferguson, & E. Moravcsik (Eds.), *Universals of Human Language* (p. 61-91). Stanford: Stanford University Press.
- Greenberg, J. (1978b). Typology and Cross-Linguistic Generalizations. In J. Greenberg, C. Ferguson, & E. Moravcsik (Eds.), *Universals of Human Language* (p. 33-59). Stanford: Stanford University Press.

- Greenberg, J. (1991). Typology/universals and second language acquisition. In T. Huebner & C. Ferguson (Eds.), *Cross Currents in Second Language Acquisition and Linguistic Theory* (p. 37-46). John Benjamins.
- Greenberg, J., Ferguson, C., & Moravcsik, E. (1978). *Universals of Human Language*. Stanford: Stanford University Press.
- Greenberg, J., Osgood, C., & Jenkins, J. (1963). Memorandum Concerning Language Universals. In J. Greenberg (Ed.), *Universals of Language* (p. 255-264).
- Gries, S. (2009). What is corpus linguistics? *Linguistics and Language Compass*, 3(5), 1225-1241. doi: 10.1111/j.1749-818X.2009.00149.x
- Gries, S. (2013). *Statistics for Linguistics with R : a Practical Introduction*. Berlin/Boston: De Gruyter Mouton.
- Grimes, B. F. (1988). *Ethnologue: Languages of the world*. Dallas: Summer Institute of Linguistics.
- Guevara, E., & Scalise, S. (2009). Searching for Universals in Compounding. In S. Scalise, E. Magni, & A. Bisetto (Eds.), *Universals of Language Today* (p. 101-128). Bologna: Springer.
- Guillaume, B. (2021). Graph Matching and Graph Rewriting: GREW tools for corpus exploration, maintenance and conversion. In P. Merlo, J. Tiedemann, & R. Tsarfaty (Eds.), *EACL 2021 16th Conference of the European Chapter of the Association for Computational Linguistics* (p. 1-9).
- Guzmán Naranjo, M., & Becker, L. (2018). Quantitative word order typology with UD. In *17th International Workshop on Treebanks and Linguistic* (p. 91-104).
- Guzmán Naranjo, M., & Becker, L. (2022). Statistical bias control in typology. *Linguistic Typology*, 26(3), 605-670. doi: 10.1515/lingty-2021-0002
- Habermas, J. (1979). *Communication and the Evolution of Society*. Boston: Beacon Press.
- Haddington, P., Rostila, J., & Tuomarla, U. (2005). Concerning the synthesis between intuition-based study of norms and observation-based study of corpora. *Sky Journal of Linguistics*, 18, 357-377.
- Haig, G., Schnell, S., & Seifart, F. (2022). *Doing Corpus-Based Typology with Sopken Language Corpora*. Honolulu: Language Documentation & Conservation.
- Hammarström, H. (2008). Counting languages in dialect continua using the criterion of mutual intelligibility. *Journal of Quantitative Linguistics*, 15(1), 34-45. doi: 10.1080/09296170701794278
- Hammarström, H. (2009). Sampling and genealogical coverage in WALs. *Linguistic Typology*, 13(1), 105-119. doi: 10.1515/LITY.2009.006
- Hammarström, H., & Donohue, M. (2014). Some Principles on the Use of Macro-Areas in

-
- Typological Comparison. *Language Dynamics and Change*, 4(1), 167-187. doi: 10.1163/22105832-00401001
- Hammarström, H., Forkel, R., Haspelmath, M., & Bank, S. (2023). *Glottolog 4.8*. Retrieved from <https://glottolog.org/>
- Harris, A. (2010). Explaining typologically unusual structures: the role of probability. In J. Wohlge-muth & M. Cysouw (Eds.), *Rethinking Universals: How Rarities affect Linguistic Theory* (p. 91-104). Berlin: De Gruyter Mouton.
- Haspelmath, M. (2009). An Empirical Test of the Agglutination Hypothesis. In S. Scalise, E. Magni, & A. Bisetto (Eds.), *Universals of Language Today* (p. 13-30). Bologna: Springer.
- Haspelmath, M. (2010). Comparative concepts and descriptive categories in crosslinguistic studies. *Language*, 86(3), 663-687.
- Haspelmath, M., Dryer, M., Gil, D., & Comrie, B. (2005). *The world atlas of language structures*. Oxford: Oxford University Press.
- Hawkins, J. (1979). Implicational Universals as Predictors of Word Order Change. *Language*, 55(3), 618-648.
- Hawkins, J. (1980). On Implicational and Distributional Universals of Word Order. *Journal of Linguistics*, 16(2), 193-235.
- Hawkins, J. (1983). *Word Order Universals*. New York: Academic Press.
- Hawkins, J. (2004). *Efficiency and complexity in grammars*. Oxford: Oxford University Press.
- Hawkins, J. (2009). Languages Universals and the Performance-Grammar Correspondence Hypoth-esis. In M. Christiansen, C. Collins, & S. Edelman (Eds.), *Language Universals* (p. 54-78). New York: Oxford University Press.
- Hawkins, J. (2010). Processing Efficiency and Complexity in Typological Patterns. In J. J. Song (Ed.), *The Oxford Handbook of Linguistic Typology* (p. 1-16). Oxford: Oxford University Press.
- Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Brain Sciences*, 33(2-3), 61-83.
- Hetterle, K. (2015). *Adverbial clauses in cross-linguistic perspective*. Berlin: De Gruyter Mouton.
- Hierro Sánchez Pescador, J. (1989). Los universales lingüísticos. In *Principios de Filosofía del Lenguaje* (p. 72-79). Spain: Alianza Editorial.
- Hill, B., Ide, S., Ikuta, S., Kawasaki, A., & Ogino, T. (1986). Universals of linguistic politeness: Quantitative evidence from Japanese and American english. *Journal of Pragmatics*, 10(3), 347-371. doi: 10.1016/0378-2166(86)90006-8
- Hinds, J. (1988). *Japanese: Descriptive Grammars*. London: Routledge.

-
- Hirst, D., & Di Cristo, A. (1998). A survey of intonation systems. *Intonation systems: A survey of twenty languages*(May), 1-44.
- Hjelmlev, L. (1963). *Language: An Introduction*. Madison: University of Wisconsin Press.
- Hockett, C. (1963). The Problem of Universals in Language. In J. Greenberg (Ed.), *Universals of Language* (p. 1-22). Cambridge, Massachusetts: The M.I.T. Press.
- Hoenigswald, H. (1963). Are There Universals of Linguistic Change? In J. Greenberg (Ed.), *Universals of Language* (p. 23-41). Cambridge, Massachusetts: The M.I.T. Press.
- Hopper, P. J., & Thompson, S. A. (1993). Language universals, discourse pragmatics, and semantics. *Language Sciences*, 15(4), 357-376. doi: 10.1016/0388-0001(93)90009-H
- Hornstein, N., & Boeckx, C. (2009). Approaching Universals from Below: I-Universals in Light of a Minimalist Program for Linguistic Theory. In M. Christiansen, C. Collins, & S. Edelman (Eds.), *Language Universals* (p. 79-98). New York: Oxford University Press.
- Hurford, J. R. (2009). Universals and the Diachronic Life Cycle of Languages. In M. Christiansen, C. Collins, & S. Edelman (Eds.), *Language Universals* (p. 40-53). New York: Oxford University Press.
- Hyman, L. (2008). Universals in phonology. *The linguistic Review*(25), 83-137. doi: 10.1515/TLIR.2008.003
- Ibbotson, P., & Tomasello, M. (2016). Evidence rebuts Chomsky's theory of language learning. *Scientific American*, 315(5), 70-75. doi: 10.1016/j.neurenf.2008.09.005
- Iosad, P. (2010). Right at the left edge: initial consonant mutations in the languages of the world. In J. Wohlgemuth & M. Cysouw (Eds.), *Rethinking Universals: How Rarities affect Linguistic Theory* (p. 105-138). Berlin: De Gruyter Mouton.
- Itkonen, E. (1980a). Qualitative vs. Quantitative Analysis in Linguistics. In T. Perry (Ed.), *Evidence and Argumentation in Linguistics* (p. 334-366). Berlin/New York: Mouton de Gruyter.
- Itkonen, E. (1980b). Reply to Dahl. In T. Perry (Ed.), *Evidence and Argumentation in Linguistics* (p. 146-162). Berlin/New York: Walter de Gruyter.
- Jakobson, R. (1958). Typological Studies and their Contribution to Historical Comparative Linguistics. In E. Siversten (Ed.), *Proceedings of the Eighth International Congress of Linguists*. Oslo.
- Jakobson, R. (1963). Implications of Language Universals for Linguistics. In J. Greenberg (Ed.), *Universals of Language* (p. 208-219). Cambridge, Massachusetts: The M.I.T. Press.
- Jameson, K. A., & Webster, M. A. (2019). Color and culture: Innovations and insights since Basic Color Terms, Their Universality and Evolution (1969). *Color Research and Application*,

- 44(6), 1034-1041. doi: 10.1002/col.22438
- Jancso, A., Moran, S., & Stoll, S. (2020). The ACQDIV Corpus Database and Aggregation Pipeline. In N. Calzolari et al. (Eds.), *Proceedings of the 12th Conference on Language Resources and Evaluation* (p. 156-165).
- Kahane, S., Caron, B., Strickland, E., & Gerdes, K. (2021). Annotation guidelines of UD and SUD treebanks for spoken corpora: a proposal. In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories*.
- Kahane, S., Peng, Z., & Gerdes, K. (2023). Word order flexibility: a typometric study. In O. Rambow & F. Lareau (Eds.), *Proceedings of the Seventh International Conference on Dependency Linguistics (Depling, GURT/SyntaxFest 2023)* (p. 1-12).
- Kahane, S., Vanhove, M., Ziane, R., & Guillaume, B. (2021). A morph-based and a word-based treebank for Beja. In D. Dakota, K. Evang, & S. Kubler (Eds.), *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories* (p. 1-13).
- Kamholz, D., Pool, J., & Colowick, S. M. (2014). PanLex: Building a resource for panlingual lexical translation. In *Proceedings of the 9th international conference on language resources and evaluation, lrec 2014* (p. 3145-3150).
- Kasper, G. (1990). Linguistic politeness: Current research issues. *Journal of Pragmatics*, 14(2), 193-218. doi: 10.1016/0378-2166(90)90080-W
- Kasper, G. (1992). Pragmatic transfer. *University of Hawaii Working papers in ESL*, 11(1), 1-34. doi: 10.1515/9783110431056-013
- Kasper, G., & Blum-Kulka, S. (1993). *Interlanguage Pragmatics*. New York: Oxford University Press. doi: 10.4324/9781315668925-14
- Kasper, G., & Dahl, M. (1991). Research Methods In Interlanguage Pragmatics. *Studies in Second Language Acquisition*, 13(2), 215-247. doi: 10.1017/S0272263100009955
- Keefe, R. (2012). Modelling vagueness: What can we ignore? *Philosophical Studies*, 161(3), 453-470. doi: 10.1007/s11098-011-9750-1
- Kemeny, J. C. (1959). *A philosopher looks at science*. New York: Van Nostrand.
- Kilgariff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovár, V., Michelfeit, J., ... Suchomel, V. (2014). The Sketch Engine: Ten Years On. *Lexicography*, 1, 7-36.
- Kiparsky, P. (2004). *Universals constrain change; change results in typological generalizations*. Stanford.
- Kirbi, J., & Gooyabadi, M. (2021). A Bayesian nonparametric mixture model for studying universal patterns in color naming. *Applied Mathematics and Computation*, 395. doi: 10.1016/j.amc.2020.125868

- Kirkegaard, E. O. W. (2016). Language typology and statistics. *Open Science Framework*. Retrieved from osf.io/d7tzt
- Kirov, C., Cotterell, R., Sylak-Glassman, J., Walther, G., Vylomova, E., Xia, P., ... Hulden, M. (2018). UniMorph 2.0: Universal Morphology. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (p. 1868-1873).
- Koptjevskaja Tamm, M. (2010). Linguistic Typology and Language Contact. In J. J. Song (Ed.), *The Oxford Handbook of Linguistic Typology* (p. 1-15). Oxford: Oxford University Press.
- Krámský, J. (1959). A Quantitative Typology of Languages. *Language and speech*, 2(2), 72-85.
- Kryk, B. (1990). On Pragmatic Universals. *Studia Anglica Posnaniensia*, 23, 63-72.
- Lahiri, A., & Plank, F. (2009). What Linguistic Universals Can be True of. In S. Scalise, E. Magni, & A. Bisetto (Eds.), *Universals of Language Today* (p. 31-58). Bologna: Springer.
- Lakoff, G. (1973). Hedges: A Study in Meaning Criteria and the Logic of Fuzzy Concepts. *Journal of Philosophical Logic*, 2(4), 183-228.
- Lappin, S., & Shieber, S. M. (2007). *Machine learning theory and practice as a source of insight into universal grammar* (Vol. 43) (No. 2).
- Leech, G. (1989). *Principles of Pragmatics*. New York: Longman.
- Leech, G., & Weisser, M. (2003). Generic Speech Act Annotation for Task-Oriented Dialogue. *Corpus Linguistics 2003*, 16, 441-446.
- Lehmann, W. P. (1978). *Syntactic typology: studies in the phenomenology of language*. Austin and London: University of Texas Press.
- Levinson, S. (2000). Yeli Dnye and the Theory of Basic Color Terms. *Journal of Linguistic Anthropology*, 10(1), 3-55. doi: 10.1525/jlin.2000.10.1.3
- Levinson, S. (2011). Pragmatics, Universals in. In P. C. Hogan (Ed.), *The Cambridge Encyclopedia of the Language Sciences* (p. 654-657). New York: Cambridge University Press.
- Levshina, N. (2019). Token-based typology and word order entropy: A study based on Universal Dependencies. *Linguistic Typology*, 23(3), 533-572. doi: 10.1515/lingty-2019-0025
- Levshina, N. (2022). Corpus-based typology: Applications, challenges and some solutions. *Linguistic Typology*, 26(1), 129-160. doi: 10.1515/lingty-2020-0118
- Levshina, N., Namboodiripad, S., Allasonnière-Tang, M., Kramer, M. A., Talamo, L., Verkerk, A., ... Liu, Z. (2021). Why we need a gradient approach to word order. *Linguistics*, 61.
- Lewin, K. (1935). *A dynamic theory of personality*. New York: McGraw-Hill.
- Lieb, H.-H. (1975). Universals of Language: Quandaries and Prospects. *Foundations of Language*, 12(4), 471-511.
- Lin, F. Y. (2017). A refutation of universal grammar. *Lingua*, 193, 1-22.

-
- Liu, H. (2008). A quantitative study of Chinese structures based on dependency treebanks. *Changjiang Xueshu*, 3, 120-128.
- Liu, H. (2010). Dependency direction as a means of word-order typology: A method based on dependency treebanks. *Lingua*, 120(6), 1567-1578. doi: 10.1016/j.lingua.2009.10.001
- Loporcaro, M. (2011). Two euroversals in a global perspective: auxiliation and alignment. In P. Siemund (Ed.), *Linguistic Universals and Language Variation* (p. 55-91). Berlin: De Gruyter Mouton.
- Luo, C. (1994). *The accessibility hierarchy and clefting* (Unpublished doctoral dissertation). University of Manitoba.
- Luotolahti, J., Kanerva, J., Pyysalo, S., & Ginter, F. (2015). SETS: Scalable and Efficient Tree Search in Dependency Graphs. In M. Gerber, C. Havasi, & F. Lacatusu (Eds.), *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations* (p. 51-55).
- Machálek, T. (2020). KonText: Advanced and Flexible Corpus Query Interface. In N. Calzolari et al. (Eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference* (p. 7005-7010).
- Maddieson, I. (2010). Typology of Phonological Systems. In J. J. Song (Ed.), *The Oxford Handbook of Linguistic Typology* (p. 1-10). Oxford: Oxford University Press.
- Magni, E. (2009). The Evloution of Latin Word (Dis)order. In S. Scalise, E. Magni, & A. Bisetto (Eds.), *Universals of Language Today* (p. 225-252). Bologna: Springer.
- Mairal, R., & Gil, J. (2004). *En torno a los universales lingüísticos*. Madrid: Cambridge University Press - AKAL.
- Majid, A., & Levinson, S. (2010). WEIRD languages have misled us, too. *Behavioral and Brain Sciences*, 33, 103. doi: 10.1017/S0140525X1000018X
- Malchukov, A. (2010). “Quirky” case: rare phenomena in case-marking and their implications for a theory of typological distributions. In J. Wohlgemuth & M. Cysouw (Eds.), *Rethinking Universals: How Rarities affect Linguistic Theory* (p. 139-168). Berlin: De Gruyter Mouton.
- Márquez Reiter, R., & Luke, K.-w. (2010). Telephone conversation openings across languages, cultures and settings. In A. Trosborg (Ed.), *Pragmatics across Languages and Cultures* (p. 103-138). Berlin: De Gruyter Mouton.
- Martí, M. A., Taulé, M., Bertran, M., & Màrquez, L. (2006). *AnCora: Multilingual and Multilevel Annotated Corpora*. Barcelona: Catalan Government.
- Masini, F., & Mattioli, S. (2019). Come fare tipologia con categorie non tradizionali? In C. Gianollo & C. Mauri (Eds.), *CLUB Working Papers on Linguistics* (p. 282-295). Bologna:

- Circolo Linguistico dell'Università di Bologna.
- Matthews, P. H. (2003). *Concise Dictionary of Linguistics*. Oxford: Oxford University Press.
- Mattiola, S. (2020). *Two language samples for maximizing linguistic variety*. Retrieved from <http://www.lilec.it/list/universalist/>
- Mauri, C. (2022a). Le parti del discorso. In N. Grandi & C. Mauri (Eds.), *La tipologia linguistica: unità e diversità nelle lingue del mondo* (p. 218-252). Roma: Carocci Editore.
- Mauri, C. (2022b). Obiettivi, metodi e strumenti della tipologia. In N. Grandi & C. Mauri (Eds.), *La tipologia linguistica: unità e diversità nelle lingue del mondo* (p. 23-54). Roma: Carocci Editore.
- Max Planck Institute for Evolutionary Anthropology. (2015). *Typological tools for field linguistics*.
- McCarthy, A. D., Wu, W., Mueller, A., Watson, B., & Yarowsky, D. (2020). Modeling color terminology across thousands of languages. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 2241-2250. doi: 10.18653/v1/d19-1229
- Meadwell, H. (1994). The foundations of Habermas's universal pragmatics. *Theory and Society*, 23(5), 711-727. doi: 10.1007/BF00992908
- Mel'cuk, I. A. (1988). *Dependency Syntax: Theory and Practice*. Albany: State University Press of New York.
- Mendívil-Giró, J.-L. (2018). Is Universal Grammar ready for retirement? A short review of a longstanding misinterpretation. *Journal of Linguistics*, 54(4), 1-25. doi: 10.1017/S0022226718000166
- Merlo, P. (2015). Predicting Word Order Universals. *Journal of Language Modelling*, 3(2), 317-344. doi: 10.15398/jlm.v3i2.112
- Merlo, P. (2016). Quantitative computational syntax: some initial results. *Italian Journal of Computational Linguistics*, 2(1), 11-29. doi: 10.4000/ijcol.347
- Mi Oh, Y., & Pellegrino, F. (2022). Towards robust complexity indices in linguistic typology. *Studies in Language*, 46(4), 1-41. doi: 10.1075/sl.22034.oh
- Miestamo, M. (2003). *Clausal negation: A typological study* (Unpublished doctoral dissertation). Helsingin yliopisto, Helsinki.
- Miestamo, M. (2005). *Standard negation: The negation of declarative verbal main clauses in a typological perspective*. Berlin: Mouton de Gruyter.
- Miestamo, M., Bakker, D., & Arppe, A. (2016). Sampling for variety. *Linguistic Typology*, 20(2), 233-296. doi: 10.1515/lingty-2016-0006

-
- Mithun, M. (1992). Is basic word order universal? In D. Payne (Ed.), *Pragmatics of Word Order Flexibility* (p. 15-59). Amsterdam/Philadelphia: John Benjamin Publishing. doi: 10.1075/tsl.22.02mit
- Moldovan, A. M., Skorvid, S. S., & Kibrik, A. A. (2004). *Jazyki mira*. Moscow: Academia.
- Moravcsik, E. (1978). Universals of Language Contact. In J. Greenberg, C. Ferguson, & E. Moravcsik (Eds.), *Universals of Human Language* (p. 93-122). Stanford: Stanford University Press.
- Moravcsik, E. A. (2010). Explaining Language Universals. In J. J. Song (Ed.), *The Oxford Handbook of Linguistic Typology* (p. 69-89). Oxford: Oxford University Press. doi: 10.1093/oxfordhb/9780199281251.013.0005
- Moravcsik, E. A. (2012). *Introducing Language Typology*. Cambridge: Cambridge University Press. doi: 10.1017/cbo9780511978876
- Moreno Cárdenas, E. L., Zapata-Zapata, A. D., & Kim, D. (2020). Modeling dark fermentation of coffee mucilage wastes for hydrogen production: Artificial neural network model vs. fuzzy logic model. *Energies*, 13, 1-13.
- Morteza, A., Seweryn, P., & Ludvine, R. (2019). *Testing Greenberg ‘s Linguistic Universals on the Universal Dependencies Corpora using a Graph Rewriting tool* (Unpublished doctoral dissertation). Université de Lorraine.
- Müller, R.-A. (2009). Languages Universals in the Brain: How Linguistic Are They? In M. Christiansen, C. Collins, & S. Edelman (Eds.), *Language Universals* (p. 224-252). New York: Oxford University Press.
- Muysken, P. (2000). *Bilingual Speech: A Typology of Code-Mixing*. Cambridge: Cambridge University Press.
- Naseem, T., Chen, H., Barzilay, R., & Johnson, M. (2010). Using universal linguistic knowledge to guide grammar induction. In H. Li & L. Marquez (Eds.), *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (p. 1234-1244).
- Newmeyer, F. J. (2007). *Possible and Probable Languages: A Generative Perspective on Linguistic Typology*. Oxford: Oxford University Press.
- Newmeyer, F. J. (2008). Universals in syntax. *Linguistic Review*, 25(1-2), 35-82. doi: 10.1515/TLIR.2008.002
- Nichols, J. (1986). Head-Marking and Dependent-Marking Grammar. *Language*, 66, 56-119.
- Nivre, J. (2015). Towards a Universal Grammar for Natural Language Processing. In A. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing* (p. 3-16). doi: 10.1007/978-3-319-18111-0
- Nivre, J., De Marneffe, M. C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C. D., ... Zeman, D.

- (2016). Universal dependencies v1: A multilingual treebank collection. *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*, 1659-1666.
- Nivre, J., De Marneffe, M. C., Ginter, F., Haji, J., Manning, C. D., Pyysalo, S., ... Zeman, F. T. D. (2020). Universal dependencies v2: An evergrowing multilingual treebank collection. In N. Calzolari et al. (Eds.), *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings* (p. 4034-4043).
- Nivre, J., De Marneffe, M.-C., Ginter, F., Hajič, J., Manning, C. D., Pyysalo, S., ... Zeman, D. (2023). *Universal Dependencies*. Retrieved from <https://universaldependencies.org/>
- Novák, V. (2008a). A comprehensive theory of thirchotomous evaluative linguistic expressions. *Fuzzy Sets and Systems*, 22(159), 2939-2969.
- Novák, V. (2008b). Mathematical fuzzy logic: From vagueness to commonsense reasoning. In G. Kreuzbauer, N. Gratzl, & E. Hielb (Eds.), *Retorische Wissenschaft: Rede und Argumentation in Theorie und Praxis* (p. 191-223). Wien: LIT-Verlag.
- Novak, V. (2021). Intermediate quantifiers in partial fuzzy type theory. *Journal of Applied Logics*, 8(5), 1215-1240.
- Novák, V., & Murinová, P. (2019). A Formal Model of the Intermediate Quantifiers “A Few”, “Several” and “A Little”. In R. Baker Kearfott, I. Batyrshin, M. Reformat, M. Ceberio, & V. Kreinovich (Eds.), *Fuzzy Techniques: Theory and Applications* (p. 429-441). doi: 10.1007/978-3-030-21920-8_39
- Novák, V., Murinová, P., & Boffa, S. (2020). On the Properties of Intermediate Quantifiers and the Quantifier “MORE-THAN”. In M. J. Lesot et al. (Eds.), *Communications in Computer and Information Science* (p. 159-172). Springer. doi: 10.1007/978-3-030-50153-2_12
- Novák, V., Perfilieva, I., & Dvorak, A. (2016). *Insight into Fuzzy Modeling*. New Jersey: Wiley & Sons.
- Novák, V., Perfilieva, I., & Mockor, J. (1999). *Mathematical principles of fuzzy logic*. Dordrecht: Luwer Academic.
- Odden, D. (2003). Languages and Universals. *Journal of Universal Language*, 4(1), 33-74. doi: 10.22425/jul.2003.4.1.33
- O’Horan, H., Berzak, Y., Vulić, I., Reichart, R., & Korhonen, A. (2016). Survey on the use of typological information in natural language processing. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 1297-1308.
- O’Keeffe, A., & McCarthy, M. (2013). *The Routledge Handbook of Corpus Linguistics*. London / New York: Routledge.
- Osgood, C. E. (1963). Language Universals and Psycholinguistics. In J. Greenberg (Ed.), *Univer-*

- sals of Language* (p. 236-254). Cambridge, Massachussetts: The M.I.T. Press.
- Ozeki, H., & Shirai, Y. (2007). Does the noun phrase accessibility hierarchy predict the difficulty order in the acquisition of Japanese relative clauses? *Studies in Second Language Acquisition*, 29(2), 169-196. doi: 10.1017/S0272263107070106
- Pagel, M. (2009). The History, Rate and Pattern of World Linguistic Evolution. In C. Knight, M. Studdert-Kennedy, & J. Hurford (Eds.), *The Evolutionary Emergence of Language* (p. 391-416). Cambridge: Cambridge University Press. doi: 10.1017/cbo9780511606441.023
- Pérez Hernández, L. (1999). Grounding politeness. *Journal of English Studies*, 1, 209-236. doi: 10.18172/jes.53
- Pericliev, V. (2012). Formulating and misformulating language universals. *Folia Linguistica*, 46(1), 211-232. doi: 10.1515/flin.2012.7
- Pericliev, V. (2018). Language Universals: Their Discovery and Verbalization by Machine. *Journal of Universal Language*, 19(1), 33-55. doi: 10.22425/jul.2018.19.1.33
- Perkins, R. D. (1989). Statistical Techniques for Determining Language Sample Size. *Studies in Language*, 13(2), 293-315. doi: 10.1075/sl.13.2.04per
- Perrier, G., & Guillaume, B. (2019). GREW, un outil au service de l'annotation de corpus et de l'exploitation de corpus annotés. In E. Schang et al. (Eds.), *Journées scientifiques Linguistique informatique, formelle de terrain* (p. 1-5).
- Petrov, S., Das, D., & McDonald, R. (2012). A Universal Part-of-Speech Tagset. In N. Calzolari et al. (Eds.), *Proceedings of the Conference on Language Resources and Evaluation* (p. 2089-2096).
- Piantadosi, S. T., & Gibson, E. (2014). Quantitative standards for absolute linguistic universals. *Cognitive Science*, 38(4), 736-756. doi: 10.1111/cogs.12088
- Plank, F. (2009). WAL values evaluated. *Linguistic Typology*, 13(1), 41-75. doi: 10.1515/LITY.2009.003
- Plank, F., & Filimonova, E. (2000). The Universals Archive: A Brief Introduction for Prospective Users. *STUF - Language Typology and Universals*, 53(1), 109-123. doi: 10.1524/stuf.2000.53.1.109
- Polinsky, M. (2010). Linguistic Typology and Formal Grammar. In J. J. Song (Ed.), *The Oxford Handbook of Linguistic Typology* (p. 1-12). Oxford: Oxford University Press.
- Polyakov, V. N., Solovyev, V. D., Wichmann, S., & Belyaev, O. (2009). Using WAL and Jazyki mira. *Linguistic Typology*, 13(1), 137-167. doi: 10.1515/LITY.2009.008
- Ponti, E. M., O'Horan, H., Berzak, Y., Vulić, I., Reichart, R., Poibeau, T., ... Korhonen, A. (2019). Modeling Language Variation and Universals: A Survey on Typological Linguistics

- for Natural Language Processing. *Computational Linguistics*, 45(3), 1-56. doi: 10.1162/coli_a_00357
- Pym, N. (1979). *Papers on Iwaidja Phonology and grammar*. Darwin: Summer Institute of Linguistics.
- Rama Martínez, E. (1996). Delimiting Semantics and Pragmatics. A Functional, Structural Criterion. *ATLANTIS*, 18(1-2), 359-367.
- Ramat, A. (2009). Typological Universals and Second Language Acquisition. In S. Scalise, E. Magni, & A. Bisetto (Eds.), *Universals of Language Today* (p. 253-272). Berlin: Springer. doi: 10.1007/978-1-4020-8825-4_13
- Ramat, P. (2009). How Universal are Linguistic Categories. In S. Scalise, E. Magni, & A. Bisetto (Eds.), *Universals of Language Today* (p. 1-12). Bologna: Springer.
- Ramat, P. (2010). The (Early) History of Linguistic Typology. In J. J. Song (Ed.), *The Oxford Handbook of Linguistic Typology* (p. 1-10). Oxford: Oxford University Press.
- Rijkhoff, J., & Bakker, D. (1998). Language Sampling. *Linguistic Typology*, 2, 263-314. doi: 10.1093/oxfordhb/9780199281251.013.0007
- Rijkhoff, J., Bakker, D., Hengeveld, K., & Kahrel, P. (1993). A Method of Language Sampling. *Studies in Language*, 17(1), 169-203. doi: 10.1075/sl.17.1.07rij
- Roberson, D., Davidoff, J., & Davies, I. (2000). Color categories are not universal: Replications and new evidence from a stone-age culture. *Journal of Experimental Psychology: General*, 129(3), 369-398. doi: 10.1037/0096-3445.129.3.369
- Romero-Trillo, J. (2008). *Pragmatics and Corpus Linguistics*. Berlin/New York: De Gruyter Mouton.
- Roque, A., Tsuetaki, A., Sarathy, V., & Scheutz, M. (2020). Developing a corpus of indirect speech act schemas. In N. Calzolari et al. (Eds.), *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings* (p. 220-228).
- Rose, K. R. (1992). Speech acts and questionnaires: The effect of hearer response. *Journal of Pragmatics*, 17(1), 49-62. doi: 10.1016/0378-2166(92)90028-A
- Rosén, V., De Smedt, K., Meurer, P., & Dyvik, H. (2012). An Open Infrastructure for Advanced Treebanking. In *META-RESEARCH Workshop on Advanced Treebanking at LREC2012* (p. 22-29).
- Rosén, V., Dyvik, H., Meurer, P., & De Smedt, K. (2017). Exploring Treebanks with INESS Search. In J. Tiedemann & N. Tahmasebi (Eds.), *Proceedings of the 21st Nordic Conference on Computational Linguistics* (p. 326-329).
- Rühlemann, C. (2011). Corpus-based pragmatics II: Quantitative studies. In W. Bublitz &

-
- N. R. Norrick (Eds.), *Foundations of Pragmatics* (p. 629-656). Berlin: De Gruyter Mouton.
 doi: 10.1515/9783110214260.629
- Ruhlen, M. (1987). *A Guide to the World's Languages* (Vol. 1). Stanford: Stanford University Press.
- Ruhlen, M. (1991). *A Guide to the World's Languages*. London: Arnold.
- Sainsbury, R. (1990). Concepts without Boundaries. In R. Keefe & P. Smith (Eds.), *Inaugural Lecture* (p. 251-264). London: King's College London Department of Philosophy.
- Sanford, D. (1980). Armstrong's theory of universals. *The British Journal for the Philosophy of Science*, 69-79.
- Sansò, A. (2022). Gli universali linguistici. In N. Grandi & C. Mauri (Eds.), *La tipologia linguistica: unità e diversità nelle lingue del mondo* (p. 55-96). Roma: Carocci Editore.
- Sapir, E. (1921). *Language: An Introduction to the Study of Speech*. New York: Harcourt Brace.
- Saporta, S. (1963). Phoneme Distribution and Language Universals. In J. Greenberg (Ed.), *Universals of Language* (p. 48-57). Cambridge, Massachusetts: The M.I.T. Press.
- Saunders, B. (1995). Disinterring Basic Color Terms: A study in the mystique of cognitivism. *History of the Human Sciences*, 8(4), 19-38. doi: 10.1177/095269519500800402
- Scalise, S., Magni, E., & Bisetto, A. (2009). *Universals of Language Today*. Netherlands: Springer.
- Schaub, W. (1985). *Babungu: Descriptive Grammars*. London: Croom Helm.
- Schnell, S., Haig, G., & Seifart, F. (2022). The role of language documentation in corpus-based typology. *Zurich Open Repository and Archive*, 1-29. doi: 10.5167/uzh-217097
- Schnell, S., & Schiborr, N. N. (2022). Crosslinguistic Corpus Studies in Linguistic Typology. *Annual Review of Linguistics*, 8, 171-191. doi: 10.1146/annurev-linguistics-031120-104629
- Schuster, S., & Manning, C. D. (2016). Enhanced English Universal Dependencies: An Improved Representation for Natural Language Understanding Tasks. In N. Calzolari et al. (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (p. 2371-2378).
- Searle, J. (1976). A classification of illocutionary Acts. *Language in Society*, 5(1), 1-23.
- Searle, J. (1994). *Actos de habla :ensayo de filosofía del lenguaje* (L. M. Valdés Villanueva, Trans.). Barcelona: Planeta-De Agostini.
- Shapiro, S. (2006). *Vagueness in Context*. Oxford: Oxford University Press.
- Siemund, P. (2011). *Linguistic Universals and Language Variation*. Berlin: De Gruyter Mouton.
- Siewierska, A. (1988). *Word Order Rules*. New York: Croom Helm. doi: 10.2307/415269
- Sigurd, B. (1989). A Supplementary Relativization Hierarchy Based on the Complexity of the Relative Phrase. *Studia Linguistica*, 43(1), 33-46. doi: 10.1111/j.1467-9582.1989.tb00793.x

- Skirgård, H., Haynie, H. J., Blasi, D. E., Hammarström, H., Collins, J., Latache, J. J., ... Gray, R. D. (2023). Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. *Science Advances*, 9.
- Smith, N. (2008). *Vagueness and degrees of truth*. Oxford: Oxford University Press.
- Smith, N., & Tsimpli, I.-M. (1995). *The mind of a savant: Language learning and modularity*. Blackwell Publishing.
- Song, J. J. (2009). Word order patterns and principles: An overview. *Linguistics and Language Compass*, 3(5), 1328-1341. doi: 10.1111/j.1749-818X.2009.00153.x
- Song, J. J. (2010). Word Order Typology. In J. J. Song (Ed.), *The Oxford Handbook of Linguistic Typology* (p. 1-21). Oxford: Oxford University Press.
- Spencer, A. (2010). Morphological Universals. In R. Mairal & J. Gil (Eds.), *Linguistic Universals* (p. 101-129). Cambridge: Cambridge University Press.
- Sridhar, S. R. (1990). *Kannada: Descriptive Grammars*. London: Routledge.
- Stabler, E. P. (2009). Computational Models of Language Universals: Expressiveness, Learnability, and Consequences. In M. Christiansen, C. Collins, & S. Edelman (Eds.), *Language Universals* (p. 1-21). Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780195305432.003.0010
- Stassen, L. (2010). The Problem of Cross-Linguistic Identification. In J. J. Song (Ed.), *The Oxford Handbook of Linguistic Typology* (p. 1-9). Oxford: Oxford University Press.
- Steedman, M. (2009). Foundations of Universal Grammar in Planned Action. In M. Christiansen, C. Collins, & S. Edelman (Eds.), *Language Universals* (p. 174-199). New York: Oxford University Press.
- Steinvall, A. (2002). *English Colour Terms in Context*. Umea: University of Umea.
- Štěpánek, J., & Pajas, P. (2010). Querying Diverse Treebanks in a Uniform Way. In N. Calzolari et al. (Eds.), *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)* (p. 1-8).
- Straka, M., Hajič, J., & Straková, J. (2016). UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In N. Calzolari et al. (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (p. 4290-4297).
- Takamura, H., Nagata, R., & Kawasaki, Y. (2016). Discriminative analysis of linguistic features for typological study. In N. Calzolari et al. (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (p. 69-76).
- Tanaka-Ishii, K. (2021). *Statistical Universals of Language*. Cham, Switzerland: Springer.

-
- Tavoni, M. (1990). La linguistica rinascimentale. In G. Lepschy (Ed.), *Storia della Linguistica* (p. 169-245). Bologna: Il Mulino.
- Tesnière, L. (1959). *Éléments de syntaxe structurale*. Paris: Klincksieck.
- Tesnière, L. (1980). *Grundzuge der strukturalen Syntax* (U. Engel, Trans.). Stuttgart: Klett-Cotta (Original work published in 1959).
- Tesnière, L. (1988). *Osnovy strukturnogo sintaksisa* (V. G. Gaka, Trans.). Moscow: Progress (Original work published in 1959).
- Tesnière, L. (1994). *Elementos de sintaxis estructural* (E. Diamante, Trans.). Madrid: Gredos (Original work published in 1959).
- Tesnière, L. (2007). *Kōzō tōgoron yōsetsu* (T. Koizumi, Trans.). Tokyo: Aoi kaba (Original work published in 1959).
- Tesnière, L. (2008). *Elementi di sintassi strutturale* (G. Proverbio & A. Trocini Cerrina, Trans.). Torino: Rosenberg & Sellier (Original work published in 1959).
- Tesnière, L. (2015). *Elements of Structural Syntax* (T. Osborne & S. Kahane, Trans.). Amsterdam/Philadelphia: John Benjamins Publishing Company (Original work published in 1959).
- The Association for Computational Linguistics. (2019). *Third Workshop on Universal Dependencies* (M. Candito, K. Evang, S. Oepen, & D. Seddah, Eds.). Stroudsburg: The Association for Computational Linguistics.
- Thomason, S. G. (2001). *Language Contact: An Introduction*. Edinburgh: Edinburgh University Press.
- Tomasello, M. (1999). *The Cultural Origins of Human Cognition*. Harvard: Harvard University Press.
- Tomasello, M. (2003). *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard: Harvard University Press.
- Tomita, M., & Carbonell, J. (1987). The Universal Parser Architecture for Knowledge-Based Machine Translation. In M. Kaufmann (Ed.), *IJCAI'87 Proceedings of the 10th international joint conference on Artificial Intelligence* (Vol. 2, p. 718-721).
- Tomlin, R. (1986). *Basic Word Order: Functional Principles*. London: Croom Helm.
- Torrens-Urrutia, A., Jiménez-López, M. D., & Novák, V. (2022). Fuzzy Natural Logic for Sentiment Analysis: A Proposal. In S. Rodriguez Gonzalez et al. (Eds.), *Lecture Notes in Networks and Systems: Distributed Computing and Artificial Intelligence, Volume 2* (Vol. 332, p. 64-73). doi: 10.1007/978-3-030-86887-1_6
- Torrens-Urrutia, A. (2019). *A Formal Characterization of Fuzzy Degrees of Grammaticality for*

- Natural Language*. Tarragona: Universitat Rovira i Virgili.
- Torrens-Urrutia, A., Jiménez-López, M. D., Brosa-Rodríguez, A., & Adamczyk, D. (2022). A Fuzzy Grammar for Evaluating Universality and Complexity in Natural Language. *Mathematics*, 10(15), 1-23. doi: 10.3390/math10152602
- Torrens-Urrutia, A., Jiménez-López, M. D., & Brosa-Rodríguez, A. (2021). A Fuzzy Approach to Language Universals for NLP. In *IEEE International Conference on Fuzzy Systems* (Vol. 2021-July, p. 1-7). Institute of Electrical and Electronics Engineers Inc. doi: 10.1109/FUZZ45933.2021.9494516
- Torrens-Urrutia, A., Novák, V., & Jiménez-López, M. D. (2022). Describing linguistic vagueness of evaluative expressions using fuzzy natural logic and linguistic constraints. *Mathematics*, 10(15). Retrieved from <https://www.mdpi.com/2227-7390/10/15/2760> doi: 10.3390/math10152760
- Trombetti, A. (1905). *L'unità d'origine del linguaggio*. Bologna: Libreria Treves.
- Trosborg, A. (2010). *Pragmatics across Languages and Cultures*. Berlin/New York: De Gruyter Mouton.
- Tye, M. (1994). Vagueness: Welcome to the Quicksand. *Southern Journal of Philosophy*, 33, 1-22.
- Ullmann, S. (1963). Semantic Unviersals. In J. Greenberg (Ed.), *Universals of Language* (p. 172-207). Cambridge, Massachussetts: The M.I.T Press.
- van der Auwera, J., & Gast, V. (2010). Categories and Prototypes. In *The Oxford Handbook of Linguistic Typology* (p. 1-17). Oxford: Oxford University Press.
- Vennemann, T. (1974). Analogy in generative grammar: the origin of word order. In *Proceedings of the Eleventh International Congress of Linguists* (p. 79-83). Bologna: Il mulino.
- Vennemann, T. (1976). Categorical grammar and the order of meaningful elements. In A. Juliano (Ed.), *Linguistic studies offered to Joseph Greenberg on the occasion of his sixtieth birthday* (p. 615-634). Saratoga, California: Anna Libri.
- Virk, S. M., Hammarström, H., Forsberg, M., & Wichmann, S. (2020). The DReaM Corpus: A Multilingual Annotated Corpus of Grammars for the World's Languages. In *Proceedings of the 12th Conference on Language Resources and Evaluation* (p. 878-884). Marseille.
- von der Gabelentz, G. (1869). Ideen zu einer vergleichenden Syntax. In M. Lazarus & H. Steinthal (Eds.), *Zeitschrift für Volkerpsychologie und Sprachwissenschaft* 6 (p. 376-384). Berlin: Dümmler's.
- von der Gabelentz, G. (1891). *Die Sprachwissenschaft: Ihre Aufgaben, Methoden und bisherigen Ergebnisse*. Leipzig: Tauchnitz.
- von der Gabelentz, G. (1894). Hypologie der Sprachen, eine neue Aufgabe der Linguistik. *Indoger-*

- manische Forschungen*, 4, 1-7.
- Von Fintel, K., & Matthewson, L. (2008). Universals in semantics. *The linguistic Review*(25), 139-201. doi: 10.1515/TLIR.2008.004
- von Schlegel, F. (1808). *Ueber die Sprache und Weisheit der Indier*. Heidelberg: Mohr und Zimmer.
- Wacewicz, S., & Żywiczyński, P. (2015). Language Evolution: Why Hockett's Design Features are a Non-Starter. *Biosemitotics*, 8(1), 29-46. doi: 10.1007/s12304-014-9203-2
- Weil, H. (1844). *L'ordre des mots dans les langues anciennes comparées aux langues modernes* (Unpublished doctoral dissertation). Faculté des Lettres de Paris, Paris.
- Weinreich, U. (1963). On the Semantic Structure of Language. In J. Greenberg (Ed.), *Universals of Language* (p. 114-171). Cambridge, Massachussetts: The M.I.T. Press.
- Weisser, M. (2003). SPAACy, A semi-automated tool for annotating dialogue acts. *International Journal of Corpus Linguistics*, 8(1), 63-74. doi: 10.1075/ijcl.8.1.03wei
- Whaley, L. (1997). *Introduction to typology: the unity and diversity of language*. Thousand Oaks: SAGE Publications.
- Whaley, L. (2010). Syntactic Typology. In J. J. Song (Ed.), *The Oxford Handbook of Linguistic Typology* (p. 1-21). Oxford: Oxford University Press.
- Whorf, B. L. (1956). *Language, thought, and reality*. Cambridge: The M.I.T Press.
- Wierzbicka, A. (1985). Different cultures, different languages, different speech acts. Polish vs. English. *Journal of Pragmatics*, 9, 145-178. doi: 10.1016/0378-2166(85)90023-2
- Wierzbicka, A. (1986). Human Emotions : Universal or Culture-Specific ? *American Anthropologist*, 88(3), 584-594.
- Wierzbicka, A. (1992). *Semantics, Culture, and Cognition: Universal Human Concepts in Culture-Specific Configurations*. New York / Oxford: Oxford University Press.
- Wierzbicka, A. (1994). "Cultural scripts": A semantic approach to cultural analysis and cross-cultural communication. *Pragmatics and Language Learning*, 5.
- Wierzbicka, A. (1996). *Semantics : Primes and Universals*. Oxford/New York: Oxford University Press.
- Wierzbicka, A. (2003). *Cross-Cultural Pragmatics*. Berlin: Mouton de Gruyter.
- Wierzbicka, A. (2006). *English: Meaning and Culture*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780195174748.001.0001
- Wohlgemuth, J., & Cysouw, M. (2010a). *Rara & Rarissima*. Berlin: De Gruyter Mouton. doi: 10.1515/9783110228557
- Wohlgemuth, J., & Cysouw, M. (2010b). *Rethinking Universals: How Rarities affect Linguistic*

- Theory*. Berlin: De Gruyter Mouton. doi: 10.1515/9783110220933
- Xia, F., Han, C. H., Palmer, M., & Joshi, A. (2001). Automatically extracting and comparing lexicalized grammars for different languages. In B. Nebel (Ed.), *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence* (p. 1321-1326).
- Yan, J., & Liu, H. (2023). Basic word order typology revisited: A crosslinguistic quantitative study based on UD and WALS. *Linguistics Vanguard*. doi: 10.1515/lingvan-2021-0001
- Zadeh, L. A. (1965). Fuzzy Sets. *Information and Control*, 1(8), 338-353.
- Zeman, D. (2008). Reusable Tagset Conversion Using Tagset Drivers. In N. Calzolari et al. (Eds.), *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*.
- Z'graggen, J. A. (1980). A comparative word list of the Northern Adelbert Range languages, Madang Province, Papua New Guinea. *Pacific Linguistics*, 31, 1-175.
- Zipf, G. (1935). *The psychobiology of language: an introduction to dynamic philology*. Cambridge, Mass: MIT Press.