RESTRICCIONES Y *GRADIENCE* EN EL PROCESAMIENTO DEL LENGUAJE NATURAL

M. DOLORES JIMÉNEZ-LÓPEZ
ADRIÀ TORRENS-URRUTIA
Universitat Rovira i Virgili, Tarragona

RESUMEN

En este trabajo, analizamos el uso del concepto de restricción en lingüística formal con el objetivo de poner de relieve las ventajas que los modelos basados en restricciones presentan para el desarrollo de sistemas de procesamiento del lenguaje natural. En general, se reconoce que el concepto de restricción es de gran utilidad tanto en la descripción como en el procesamiento del lenguaje natural. Son muchos los modelos gramaticales que incorporan esta noción. En este trabajo, tras analizar los distintos tipos de restricciones que aparecen en los diferentes modelos, presentamos dos formalismos que consideramos de especial interés para el procesamiento del lenguaje natural: las gramáticas de propiedades y las womb grammar.

Palabras clave: restricciones, *gradience*, gramaticalidad, análisis sintáctico.

ABSTRACT

In this paper, we analyze the use of constraints in formal linguistics with the aim of highlighting the advantages of constraints based models for the development of natural language processing systems. In general, it is recognized that the concept of restriction is useful both in the description and processing of natural language. There are many grammatical models that incorporate this notion. In this paper, after analyzing the different types of restrictions that appear in different models, we present two formalisms that we consider of particular interest for the processing of natural language: property grammars and womb grammars.

Keywords: constraints, gradience, grammaticality, parsing.

1. RESTRICCIONES EN LINGÜÍSTICA

El concepto de restricción proviene de las ciencias de la computación. En lingüística, se ha utilizado sobre todo en el ámbito de la sintaxis y la fonología. Sin embargo, puede aplicarse a cualquier componente de la gramática de una lengua. En general, se reconoce que el concepto de restricción es de gran utilidad tanto en la descripción como en el procesamiento del lenguaje natural (Blache et al. 2014). Por ello, son muchas las teorías lingüísticas que utilizan restricciones en sus modelos. La noción de restricción empieza a ocupar un lugar importante en lingüística a partir de la introducción de las gramáticas de unificación.

Las teorías que utilizan esta noción entienden que las lenguas naturales poseen restricciones que definen la relación entre dos elementos. Estas restricciones señalan las propiedades que un objeto debe satisfacer. Esto es, un *input* es aceptado o rechazado en función de si satisface o viola las restricciones de una lengua. Se trata, en definitiva, de estipular propiedades que descarten/eliminen las estructuras que no pertenecen a la lengua.

Se distinguen dos tipos de restricciones: 1) las restricciones *generales* o *universales* que son válidas para cualquier lengua; y 2) las restricciones *específicas* que son aplicables a una lengua determinada.

Bajo el término genérico de gramáticas de restricciones encontramos modelos como los siguientes: *Functional Unification Grammar* (Kay 1979), *Lexical Functional Grammar* (LFG) (Kaplan y Bresnan 1982), *Categorial Grammar* (Buszkowski et al. 1988), *Head-Driven Phrase-Structure Grammar* (HPSG) (Pollard y Sag 1994), *Tree Adjoining Grammar* (TAG) (Joshi et al. 1975), *Optimality Theory* (Prince y Smolensky 1993), etc.

En la bibliografía, se establece una diferencia entre los formalismos que combinan mecanismos generativos con restricciones y los modelos basados exclusivamente en la satisfacción de restricciones. En el primer bloque, se sitúan teorías como la *HPSG*, la *LFG*, la *TAG* y la *Construction Grammar* (Goldberg 1995) entre otras. Estos modelos suelen centrarse en el uso de restricciones para definir los requerimientos argumentales entre constituyentes. De hecho, el propósito de estos modelos es *generar* estructuras y, para ello,

combinan mecanismos generativos con restricciones. Estos sistemas describen sin problemas las estructuras sintácticas en construcciones canónicas, pero ofrecen escasa información cuando un *input* presenta algún tipo de violación. Esta limitación es superada en los modelos que se basan exclusivamente en la satisfacción de restricciones, como por ejemplo, la *Optimality Theory,* las *Property Grammars* (Blache y Balfourier 2001) o las *Womb Grammars* (Dahl y Miralles 2012).

En este trabajo, presentamos las *property grammars* y las w*omb grammars,* dos formalismos relativamente recientes que, por sus características, pueden resultar especialmente adecuados para diseñar herramientas lingüísticas que den cuenta de los distintos niveles de gramaticalidad de una estructura (Aarts 2004; Keller 2000).

2. *PROPERTY GRAMMARS*

Las *property grammars,* introducidas por Blache y Balfourier en 2001, se definen como un formalismo basado exclusivamente en el uso restricciones. Blache (2005) describe así su modelo:

> It is a fully constraint based theory. In this approach, all kinds of linguistic information is [sic] represented by means of constraints. The constraint system constitutes then the core of the theory: it is the grammar, but it also constitutes, after evaluation for a given input, its description. Property Grammars is then a non-generative theory in the sense that no structure has to be build [sic], only constraints are used both to represent linguistic information and to describe inputs.

En las *property grammars,* se parte de la idea de que en toda lengua se pueden identificar una serie de regularidades que pueden ser descritas de forma independiente (las palabras presentan determinado orden; algunas palabras se excluyen mutuamente en determinado contexto; algunas palabras coocurren sistemáticamente, etc.). Se trata, simplemente, de especificar cada uno de estos tipos de información a través de distintas restricciones o propiedades. En las *property grammars* se especifican los siguientes tipos de restricción:

- **Constituencia** (*Const*): una vez identificadas todas las categorías que puedan asociarse a cada uno de los ítems léxicos, esta propiedad define las construcciones sintácticas que pueden describir a un *input*. Más que una restricción, esta

propiedad se entiende como el paso previo a la verificación (o caracterización) del resto de las propiedades.

- **Obligación** (*Oblig*): identifica los elementos imprescindibles en una construcción. Esta restricción cubre la noción de núcleo sintáctico.
- **Precedencia** (<): define el orden lineal entre dos elementos en un contexto. Ejemplo: Det < N.
- **Unicidad** (*Uniq*): hace referencia a la posibilidad de repetición de un elemento y/o categoría en una misma construcción.
- **Requerimiento** (⇒): específica que dos categorías deben ocurrir de manera conjunta en una construcción. Ejemplo: Det ⇒ N en la construcción de sujeto.
- **Exclusión** (*Excl*): estipula la no coocurrencia de dos elementos o categorías en un mismo contexto o construcción.
- **Dependencia** (*Dep*): define las relaciones de dependencia entre todos los constituyentes que figuran en una construcción.
- **Concordancia** (*Agree*): tiene en cuenta los diferentes tipos de relaciones de concordancia entre categorías.

Cabe destacar que la lista de propiedades es flexible, es decir, podemos añadir y eliminar propiedades en función del módulo lingüístico o de la lengua que se quiera describir.

A diferencia de muchos modelos de restricciones que solo tienen en cuenta una propiedad, las *property grammars* consiguen unir varios usos del concepto de restricción en un modelo único y homogéneo. Asimismo, gracias a que las restricciones son representadas independientemente, las propiedades resultan fácilmente evaluables.

En este formalismo, la información lingüística es presentada de manera autónoma, lo que permite al sistema describir cualquier tipo de *input*. Por consiguiente, con independencia del grado de gramaticalidad que presente un *input*, el sistema nos proporcionará una caracterización con una lista de las propiedades satisfechas y violadas. A diferencia de otras gramáticas de restricciones, las *property grammars* nos proporcionan mucha información acerca de las violaciones, ya que nos indican qué tipo de violación se ha producido, dónde ha tenido lugar, y por qué no se corresponde con la gramática.

La Tabla 1 recoge un ejemplo muy simple que refleja el funcionamiento de las *property grammars.*

El rojo libro	
Categorización (Conjunto inicial): {Det1 Adj2 N3} *Constituencia*: {{Det1 Adj2} {Det1 Adj2 N3}}	
Asignación	Propiedades
{Det1 Adj2}	P^+= {*Uniq* (Det); *Agree* (Det, Adj)} P^-= {Det < N}
{Det1 Adj2 N3}	P^+= {N ⇒ Det; *Uniq* (Det, N); *Oblig* (N) ; *Excl* (N, Pron); *Agree* (Det, N, Adj); *Dep* (Adj *mod* N)} P^-= {Det < N}

El libro rojo	
Categorización (Conjunto inicial): {Det1 N2 Adj3} *Constituencia*: {{Det1 N2} {Det1 N2 Adj3}}	
Asignación	Propiedades
{Det1 N2}	P^+= {Det < N; N ⇒ Det; *Uniq* (Det, N); Oblig (N) ; *Excl* (N, Pron) ; *Agree* (Det, N)} P^-= ∅
{Det1 N2 Adj3}	P^+= {Det < N; N ⇒ Det; *Uniq* (Det, N); *Oblig* (N) ; *Excl* (N, Pron) ; *Agree* (Det, N, Adj) ; *Dep* (Adj *mod* N)} P^-= ∅

Tabla 1. Ejemplo funcionamiento *property grammars*.

Este funcionamiento consiste en tres fases sucesivas:

- **Categorización**: En esta primera fase, se activan todas las categorías que puedan asociarse a cada uno de los ítems léxicos que contiene la entrada y se les asigna una posición. Se obtiene así un *conjunto de categorías*.
- **Asignación**: En la segunda fase, se identifican todas las posibles construcciones que puedan describir el input, obteniendo un *conjunto de construcciones*.

- **Caracterización**: Finalmente, en esta fase, se verifican las propiedades de cada construcción. Se obtiene de esta forma una *lista de propiedades violadas (P^-) y satisfechas (P^+).*

3. *WOMB GRAMMARS*

Las *Womb Grammars,* introducidas por Dahl y Miralles en 2012, se definen como:

> A novel constraint-based framework particularly useful for inducing, from known linguistic constraints that describe phrases in a language called the source, the linguistic constraints that describe phrases in another language, called the target. […] A formalism for inferring a language's syntax given its lexicon, a sufficiently representative set of correct phrases in it, and the property-based syntax of another language (Dahl y Miralles 2012).

Las *womb grammars* surgen como posible respuesta a la pregunta: ¿podemos inferir la sintaxis de una lengua meta a partir de la gramática de una lengua fuente? El formalismo pretende ofrecer una herramienta para generar de forma automática la gramática de lenguas desconocidas a partir de la gramática de lenguas que ya tenemos descritas.

Para conseguir inferir la sintaxis de lengua meta, el sistema requiere varios elementos: 1) el lexicón de la lengua meta; 2) un corpus representativo de frases en la lengua meta; y 3) la gramática (sintaxis) de la lengua fuente definida con gramáticas de propiedades; 4) un *parser* basado en restricciones que genere una lista de propiedades satisfechas y violadas.

Con estos elementos, las *womb grammars* aplican la gramática de propiedades que define la sintaxis de la lengua fuente al léxico y al corpus representativo de frases de la lengua meta. Este proceso de análisis, en el que se mezclan la sintaxis de una lengua ya conocida con el léxico de una lengua de la que desconocemos su sintaxis, se denomina *parser híbrido*. La aplicación de este *parser* proporciona una lista de propiedades satisfechas y violadas ya que, naturalmente, muchas de las restricciones que se cumplen en la lengua fuente son violadas en la lengua meta. Esto significa que las propiedades satisfechas son restricciones compartidas por ambas lenguas, mientras

que las restricciones violadas indican todas aquellas propiedades sintácticas que diferencian a la lengua meta de la lengua fuente. Las violaciones que se producen de manera sistemática se entienden como estructuras correctas en la lengua meta. En este punto del proceso, entra en juego el *módulo de reparación de la gramática*. Este módulo transforma la gramática de la lengua fuente en la gramática de la lengua meta, convirtiendo las restricciones recursivamente violadas en restricciones que deben ser satisfechas. De esta manera, se infiere la sintaxis (gramática) de la lengua meta.

Además del *parser* híbrido, Dahl y Miralles (2012) definen un *parser universal* que podría inferir la sintaxis de cualquier lengua a partir de una *gramática universal*. En este caso, el *parser* opera –de manera similar al caso anterior— con un corpus y un lexicón en la lengua meta; una sintaxis universal y un *womb grammar parser*. La gramática universal debería tener en cuenta todas las restricciones que puedan existir en cualquier lengua. Si esto es así, aplicado el proceso de *parsing* con el *womb grammar parser*, lo que obtendremos será una lista de propiedades violadas y una lista de propiedades satisfechas. Para obtener la gramática de nuestra lengua, bastará con descartar todas las propiedades violadas e incluir en la gramática todas las satisfechas.

4. MODELOS DE RESTRICCIONES Y *GRADIENCE*

Como hablantes, comprendemos y aceptamos oraciones con errores. Frente al procesamiento humano del lenguaje, el procesamiento automático suele estar basado en gramáticas discretas, no contempla la gradualidad en el lenguaje y no puede, por tanto, analizar estructuras no canónicas. El auge de las tecnologías lingüísticas aplicadas al análisis del lenguaje utilizado en la comunicación mediatizada por ordenador ha puesto de relieve los límites de las herramientas computacionales basadas en modelos lingüísticos que no toleran distintos niveles de gramaticalidad. Creemos que es necesaria una reformulación de los modelos gramaticales usados en los algoritmos de *parsing* convencionales. Esta reformulación exige el abandono de la concepción dicotómica de la

noción de gramaticalidad que es la que impide el análisis de toda estructura que viole mínimamente las reglas.

Los formalismos gramaticales basados en el concepto de restricción pueden ser una buena herramienta para definir modelos sintácticos que toleren distintos niveles de gramaticalidad. En concreto, defendemos que las *property grammars* pueden ser muy útiles a la hora de establecer mecanismos que permitan el análisis sintáctico de cualquier estructura sintáctica con independencia del nivel de gramaticalidad que presente (Blache y Prost 2008).

Son varias las ventajas que la utilización de las *property grammars* presenta para la definición de un modelo que dé cuenta de los niveles de gramaticalidad: las restricciones son totalmente autónomas y esto las hace potencialmente evaluables; proporcionan mucha información lingüística con independencia de que el input sea gramatical o agramatical; permiten que para la evaluación de los niveles de gramaticalidad se tenga en cuenta la importancia de la restricción violada y el valor acumulativo de violaciones. Además, el valor acumulativo de propiedades satisfechas puede equilibrar el valor final de gramaticalidad.

5. CONCLUSIONES

Son muchas las ventajas que se atribuyen a los modelos basados en restricciones. De entre todas ellas, destacamos su adecuación para diseñar herramientas lingüísticas que den cuenta de los niveles de gramaticalidad que puede presentar una estructura.

Consideramos que la introducción de la noción de *gradience* en los formalismos gramaticales es de vital importancia para la implementación de *parsers* que sean capaces de analizar estructuras no totalmente gramaticales y que calculen el nivel de gramaticalidad de una construcción. Para definir este tipo de formalismos son realmente útiles los modelos basados en restricciones ya que estos no solo analizan todas las entradas (correctas o no) sino que además proporcionan datos sobre qué restricciones son satisfechas y cuáles son violadas. Disponer de este tipo de datos nos permite realizar una estimación del grado de gramaticalidad de las distintas estructuras.

La introducción de modelos lingüísticos que definan, mediante el uso de gramáticas de restricciones, la idea de *gradience* puede tener importantes repercusiones en el procesamiento del lenguaje natural en todos aquellos ámbitos en los que las herramientas de *parsing* se enfrentan a la difícil tarea de analizar lenguaje susceptible de presentar incorrecciones o desviaciones de las reglas. El desarrollo de *parsers* para el diagnóstico de errores en la enseñanza de segundas lenguas; el diseño de herramientas para detectar el grado de deterioro lingüístico en personas que sufren patologías lingüísticas; o el análisis del denominado "*noisy text*" (Baldwin et al. 2013) en Minería de Datos Web son solo algunos de los ámbitos que podrían beneficiarse de la combinación de restricciones y *gradience* en el procesamiento del lenguaje natural.

REFERENCIAS BIBLIOGRÁFICAS

Aarts, B. 2004. "Conceptions of gradience in the history of linguistics", *Language Sciences*, 26: 343-389.

Baldwin, T., Cook, P., Lui, M., MacKinlay, A., Wang, L. 2013. "How noisy social media text? How different social media sources?" In Proceedings of 6th International Joint Conference on Natural Language Processing. Nagoya.

Blache, P. 2005. "Property Grammars: A Fully Constraint-Based Theory", *LNAI* 3438. Berlin: Springer.

Blache, P., Balfourier, J. M. 2001. "Property grammars: A flexible constraint-based approach to parsing". In *Proceedings of Seventh International Workshop on Parsing Technologies*. Beijing: Tsinghua University Press.

Blache, P., Christiansen, H., Dahl, V., Duchier, D., Villadsen, J. 2014. *Constraints and language*. Newcastle upon Tyne: Cambridge Scholars Publishing.

Blache, P., Prost, J. 2008. "A Quantification Model of Grammaticality". In *Proceedings of CSLP-08.*

Buszkowski, W., Marciszewski, W., van Benthen, J. (eds.) 1988. *Categorial Grammar*. Amsterdam: John Benjamins.

Dahl, V., Miralles, J. E. 2012. "Womb grammars: Constraint solving for grammar induction". In Sneyers, J., Frühwirth, T. (eds.).

Proceedings of the 9th Workshop on Constraint Handling Rules, Technical Report CW 624, 32-40.

Goldberg, A. 1995. *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago: Chicago University Press.

Joshi, A., Levy, L., Takahashi, M. 1975. "Tree adjunct grammars", *Journal of the Computer and System Sciences*, 10(1): 136-163.

Kaplan, R.M., Bresnan, J. 1982. "Lexical-functional grammar: A formal systems for grammatical representations". In Bresnan, J. (ed.), *The mental representation of grammatical relation*. Cambridge: MIT Press, 173-281.

Kay, M. 1979. Functional grammar. In Chiarello et al. (eds.). *Proceedings of the 5th meeting of the Berkeley Linguistics Society*. Berkeley Linguistics Society, 142-158.

Keller, F. 2000. *Gradience in grammar: experimental and computational aspects of degrees of grammaticality*. PhD thesis, University of Edinburgh.

Prince, A., Smolensky, P. 1993. *Optimality theory: Constraint interaction in generative grammar.* Technical Report, New Brunswick: Rutgers University.

Pollard, C., Sag, I. 1994. *Head-Driven Phrase Structure Grammar*. Chicago: Chicago University Press.