MODELOS COMPUTACIONALES DE LA ADQUISICIÓN Y COMPLEJIDAD LINGÜÍSTICA[1]

M. Dolores Jiménez López
Universitat Rovira i Virgili

RESUMEN

En este trabajo, proponemos la utilización de modelos computacionales de la adquisición del lenguaje para calcular la complejidad de las lenguas naturales. Presentamos un modelo de aprendizaje automático que permite calcular la complejidad relativa de las lenguas. El modelo utiliza datos reales, no requiere ningún conocimiento previo sobre la lengua y aprende de forma incremental. Defendemos que este modelo puede proporcionar datos para calcular la complejidad lingüística que serían difíciles de obtener observando el proceso de adquisición a través de la investigación experimental.

Palabras clave: complejidad, computación, adquisición del lenguaje

ABSTRACT

In this paper, we propose the use of computational models of language acquisition to calculate the complexity of natural languages. We present a machine learning model that allows us to calculate the relative complexity of languages. The model uses real data, does not require any prior knowledge about the language and learns incrementally. We argue that this model can provide data to calculate linguistic complexity that would be difficult to obtain by observing the acquisition process through experimental research.

Keywords: complexity, computation, language acquisition

[1]Este trabajo se ha realizado en el marco del proyecto FFI2015-69978-P (MINECO/FEDER, UE) financiado por el Ministerio de Economía y Competitividad y el Fondo Europeo de Desarrollo Regional.

1. INTRODUCCIÓN

El cambio que se ha producido recientemente en lingüística en lo referente a los estudios sobre la complejidad de las lenguas es considerable. Se ha pasado de negar la posibilidad de calcular la complejidad –postura defendida por gran parte de los lingüistas durante el siglo XX– a un gran interés por los estudios sobre complejidad lingüística a partir de 2001 (McWhorter, 2001).

Durante el siglo XX, imperó el dogma de la equicomplejidad en cualquiera de sus tres versiones: todas las lenguas tienen el mismo nivel de complejidad; las lenguas son inconmensurables en lo que a complejidad se refiere; la medición de la complejidad lingüística es irrelevante para el conocimiento de las lenguas y su funcionamiento. Frente a esta postura, a principios del siglo XXI, un nutrido grupo de investigadores defiende que es difícil aceptar que todas las lenguas sean iguales en su complejidad total y que la complejidad en un área de la lengua sea compensada por simplicidad en otra. Se cuestiona, por tanto, la equicomplejidad y se suceden monografías, artículos, congresos que de un modo u otro se ocupan de medir la complejidad de las lenguas (Kusters, 2003; Dahl, 2004; Miestamo et al., 2008; Sampson et al., 2009; Givon, 2009; Trudgill, 2011; McWhorter, 2012; Kortmann y Szmrecsanyi, 2012; Newmeyer y Preston, 2014; Baechler y Seiler, 2016; Baerman et al., 2015; Di Domenico, 2017; La Mantia et al., 2017; Coloma, 2017). ¿A qué se debe este cambio radical? Las razones pueden ser muy variadas, pero creemos que la falta de investigación sistemática que pruebe la supuesta equicomplejidad de las lenguas y la gran cantidad de investigaciones sobre complejidad y sistemas complejos en áreas diversas son dos motivos que han podido propiciar el auge de los estudios sobre complejidad lingüística.

A pesar de los numerosos estudios realizados, seguimos sin disponer de una solución para cuantificar la complejidad de las lenguas y cada uno de los modelos propuestos –tanto los que se ocupan de complejidad absoluta como los que abordan la complejidad relativa y con independencia del tipo de formalismo utilizado (entropía de Shannon y complejidad de Kolmogorov (Dahl, 2004; Juola, 2008; Bane, 2008; Miestamo, 2008); modelos computacionales basados en gramáticas de restricciones (Blache, 2011); teoría de sistemas complejos (Andrason, 2014), etc.)– presenta ventajas e inconvenientes.

En este trabajo, presentamos un modelo computacional que está siendo utilizado en un proyecto en el que se usan modelos de aprendizaje automático para determinar el nivel de complejidad relativa de las lenguas naturales.

2. COMPLEJIDAD LINGÜÍSTICA

A pesar del interés por los estudios sobre complejidad lingüística en los últimos años y aunque en general parece claro que las lenguas exhiben distintos niveles de complejidad, no es sencillo calcular exactamente esas diferencias. Parte de esa dificultad puede deberse a las diferentes maneras de entender la complejidad en el estudio de las lenguas naturales.

En un reciente artículo, Pallotti (2015) distingue tres significados de complejidad en los estudios lingüísticos:

- Complejidad estructural: propiedad formal de los sistemas lingüísticos relacionada con el número de elementos.
- Complejidad cognitiva: coste de procesamiento de las estructuras lingüísticas.
- Complejidad de desarrollo: orden en el que las estructuras lingüísticas emergen o son aprendidas en los procesos de adquisición y aprendizaje de primeras y segundas lenguas.

Estos tres significados se corresponden con los dos tipos básicos de complejidad que se encuentran en la bibliografía: la *complejidad absoluta* y la *complejidad relativa*. La complejidad absoluta puede ser entendida como complejidad estructural, ya que se calcula en términos de número de partes de un sistema o número de interrelaciones entre las partes (Dahl, 2004). Por su parte, la complejidad relativa incluiría tanto la complejidad cognitiva como la complejidad de desarrollo, ya que se define como el tipo de complejidad que tiene en cuenta a los usuarios del lenguaje y se mide en términos de dificultad o coste de procesamiento, aprendizaje o adquisición (Kusters, 2003).

Los estudios sobre complejidad lingüística realizados en los últimos veinte años suelen adoptar una perspectiva absoluta del concepto ya que, en general, los investigadores coinciden en que es más factible abordar la complejidad desde un punto de vista objetivo que desde un punto de vista relacionado con el usuario. Por otra parte, los estudios que adoptan esta última perspectiva prefieren ocuparse del aprendizaje de segundas lenguas debido a los problemas que los métodos observacionales y experimentales para estudiar la adquisición del lenguaje pueden plantear al estudio de la complejidad lingüística.

Teniendo en cuenta que algunas de las ideas que respaldan el dogma de la equicomplejidad se basan en el proceso de adquisición del lenguaje, creemos que los estudios sobre complejidad relativa deberían considerar este proceso para determinar las diferencias entre las lenguas naturales.

Los modelos computacionales de la adquisición pueden ser una herramienta importante para evitar los problemas de otros métodos y considerar a los niños y al proceso de adquisición como usuarios y uso adecuados para la evaluación de la complejidad lingüística.

3. MODELOS COMPUTACIONALES DE LA ADQUISICIÓN Y COMPLEJIDAD LINGÜÍSTICA

En el ámbito de los estudios sobre adquisición del lenguaje, son muchos los especialistas que reconocen que el uso de herramientas computacionales ofrece importantes ventajas metodológicas, entre las que podemos destacar las siguientes (Alishahi, 2011; Pearl, 2010):

- Alto nivel de precisión: la necesidad de explicitar todos los supuestos del modelo –tanto los que hacen referencia a los datos como aquellos que especifican el mecanismo de aprendizaje– obliga a presentar las teorías con un alto nivel de precisión. Esta propiedad distingue los modelos computacionales de las teorías lingüísticas, que normalmente no suelen ahondar en detalles, dificultando así su evaluación.
- Control sobre los datos: a diferencia de lo que ocurre en los estudios experimentales, los modelos computacionales permiten tener un control total sobre los datos, que pueden ser fácilmente manejados para observar cuáles son las consecuencias sobre el proceso de adquisición.
- Comportamiento observable: cuando se ejecuta un modelo, el impacto de cada factor en los datos de entrada o en el proceso de aprendizaje se puede estudiar directamente en su comportamiento (*output*). Esto permite que los diversos aspectos del mecanismo de aprendizaje puedan ser modificados y que sea posible estudiar los patrones de comportamiento que producen estos cambios. Por otra parte, es posible comparar el rendimiento de dos mecanismos diferentes sobre el mismo conjunto de datos, algo que es muy difícil en un estudio experimental.
- Flexibilidad: gracias a la flexibilidad de los modelos computacionales, es posible simular nuevos contextos de aprendizaje para observar su efecto sobre los mecanismos propuestos, permitiendo realizar predicciones sobre condiciones de aprendizaje que no han sido estudiadas previamente.

Además de las ventajas enumeradas, uno de los principales beneficios de los modelos computacionales es el tipo de preguntas que estos formalismos pueden responder. Si consideramos que la investigación sobre la adquisición del

lenguaje debe ocuparse de tres cuestiones diferentes (Pearl, 2010) –*qué* saben los niños, *cuándo* lo saben y *cómo* lo aprenden–, podemos afirmar que:

- la *investigación teórica* se ocupa del qué, esto es, qué conocimiento adquieren los niños durante el proceso de adquisición;
- los *análisis experimentales* proporcionan información sobre el cuándo, es decir, a qué edad el niño adquiere conocimientos lingüísticos específicos, y
- los *modelos computacionales* pueden explicar el cómo, esto es, pueden dar cuenta del proceso de adquisición del lenguaje, ya que estos modelos tienen como principal objetivo proporcionar simulaciones del mecanismo de adquisición.

Es precisamente el hecho de que los modelos computacionales de la adquisición se ocupen de explicar el proceso a través del cual se adquiere una lengua natural lo que nos lleva a proponer su uso en el ámbito de los estudios sobre complejidad lingüística.

4. APRENDIZAJE AUTOMÁTICO Y COMPLEJIDAD RELATIVA

Teniendo en cuenta la importancia de considerar el proceso de adquisición del lenguaje en los estudios sobre complejidad lingüística y siendo conscientes de las ventajas metodológicas que ofrecen los modelos computacionales de la adquisición, proponemos el uso de modelos de aprendizaje automático para determinar el nivel de complejidad relativa de las lenguas. En concreto, utilizamos modelos de inferencia gramatical, ya que estos pueden ser vistos como modelos computacionales de la adquisición dada la analogía que puede establecerse entre el proceso de adquisición y un problema de inferencia gramatical.

En todo problema de inferencia gramatical, tenemos un profesor que proporciona datos sobre el lenguaje que se quiere aprender y un aprendiz (o algoritmo de aprendizaje) que debe identificar el lenguaje subyacente a partir de los datos que recibe del profesor. Este funcionamiento guarda un paralelismo claro con el proceso de adquisición del lenguaje; en lugar de un profesor y un aprendiz, hablaríamos de un adulto y un niño que adquiere una lengua a partir de los datos que recibe.

La inferencia gramatical (de la Higuera, 2010) se incluye dentro del aprendizaje automático y se ocupa del aprendizaje de lenguajes formales. Los estudios de inferencia gramatical surgen a finales de los 60 con el intento de

Gold (1967) de formalizar la adquisición del lenguaje. Desde su nacimiento se han propuesto modelos muy diversos y, en general, podemos distinguir dos aproximaciones: una aproximación *teórica* que se centra en la obtención de resultados formales sin estudiar la relevancia lingüística de los mismos; y una aproximación más *práctica o aplicada* que intenta desarrollar sistemas que aprendan gramáticas a partir de datos reales. El modelo que utilizamos para calcular la complejidad relativa de las lenguas se incluye dentro de esta segunda aproximación y ha sido propuesto por Becerra-Bonache et al. (2015).

El modelo introducido en Becerra-Bonache et al. (2015) se presenta como un modelo de *grounded language learning* en el que se utilizan técnicas de programación lógico inductiva (*Inductive Logic Programming (ILP) tecniques*) para que el sistema aprenda un lenguaje a partir de frases que se producen en un contexto. El funcionamiento del modelo puede resumirse como sigue:

- El sistema recibe como input una base de datos formada por pares frases/imágenes (S, I), donde S es una frase relacionada con una imagen I.
- Cada imagen se transforma en una escena. Para dar cuenta de las escenas se proporciona una representación, en lógica de primer orden, que describe las propiedades de los objetos que aparecen en la imagen y las relaciones entre ellos. En ningún caso se proporcionan los posibles significados de la frase producida.
- El algoritmo debe computar los significados asociados a las frases (n-gramas) producidas para construir un modelo de lenguaje.

Básicamente, lo que hace el modelo es analizar una serie de frases y el contexto en el que estas se producen para crear un modelo de lenguaje a partir del cual poder generar lenguaje relacionando frases con significados y significados con frases.

Una importante diferencia con respecto a otros modelos es que, en este caso, el sistema aprende exclusivamente a partir de pares frase/contexto, mientras que en otros modelos se proporciona al algoritmo un conjunto de posibles significados para cada frase.

Para realizar los experimentos se ha utilizado una base de datos que contiene 10.000 imágenes con tres frases cada una (estas frases han sido generadas por hablantes reales). Para evaluar la actuación del sistema se han utilizado distintas medidas:

- precisión de los significados referenciales aprendidos;
- capacidad del sistema para comentar imágenes no mostradas, esto es, capacidad de generar frases relevantes encadenando n-gramas;
- exactitud: frases correctas que el sistema es capaz de producir;
- completitud: capacidad para producir todas las frases correctas posibles.

¿Cómo analizar los resultados obtenidos de los experimentos realizados con esta herramienta en términos de complejidad lingüística? El modelo utiliza el mismo algoritmo para aprender cualquier lengua, estableciendo así un paralelismo con la capacidad innata que permite a los humanos aprender cualquier lengua a la que sean expuestos. El modelo permite contar la cantidad de interacciones (secuencias imagen/frase) que necesita el algoritmo para llegar a un buen nivel de actuación en un dominio concreto de la lengua y muestra que, con el mismo algoritmo, el número de interacciones necesarias para adquirir un buen dominio lingüístico no es igual para todas las lenguas. Este resultado puede interpretarse en términos de complejidad: el coste/dificultad para adquirir todas las lenguas no es idéntico (ya que se requiere un número diferente de interacciones), por tanto, las lenguas varían en su complejidad relativa.

Este modelo presenta numerosas ventajas para el cálculo de la complejidad lingüística, entre las que destacamos las siguientes: utiliza datos reales y algoritmos psicológicamente plausibles; se centra en el proceso de aprendizaje; no requiere ningún conocimiento previo sobre la lengua; y aprende de forma incremental.

5. CONCLUSIONES

En este trabajo, hemos defendido que el uso de herramientas computacionales para estudiar la adquisición del lenguaje natural ofrece muchas ventajas metodológicas que pueden ser aprovechadas en los estudios sobre complejidad lingüística.

El modelo de aprendizaje automático presentado, como cualquier simulación computacional, permite realizar algunas operaciones que no podrían lle-

varse a cabo en estudios con informantes reales. Por tanto, puede proporcionar datos para calcular la complejidad lingüística que serían difíciles de obtener observando el proceso de adquisición a través de la investigación experimental. A diferencia de los experimentos psicolingüísticos con niños, este modelo evita el problema de la influencia de factores externos (no lingüísticos) que pueden condicionar el proceso de adquisición. Además, permite reproducir el mismo contexto y las mismas condiciones para la adquisición de cualquier lengua.

Los resultados de los estudios sobre complejidad lingüística pueden tener importantes implicaciones en la descripción, análisis y procesamiento del lenguaje. Los modelos computacionales de la adquisición pueden ser parte de la solución al enorme problema de calcular la complejidad de las lenguas naturales.

REFERENCIAS BIBLIOGRÁFICAS

Alishahi, A. 2011. *Computational Modeling of Human Language Acquisition.* Toronto: Morgan and Claypool.

Andrason, A. 2014. "Language complexity: An insight from complex-system theory". *International Journal of Language and Linguistics,* 2: 74-89.

Baechler, R. y Seiler, G. 2016. *Complexity, Isolation, and Variation*. Berlin: Mouton de Gruyter.

Baerman, B., Brown, D. y Corbett, G. 2015. *Understanding and Measuring Morphological Complexity*. Oxford: Oxford University Press.

Bane, M. 2008. "Quantifying and Measuring Morphological Complexity". En Ch. Chang y H. Haynie (eds.), *Proceedings of the 26th West Coast Conference on Formal Linguistics*. Somerville: Cascadilla Proceedings Project, 69-76.

Becerra-Bonache, L., Blockeel, H., Galván, M. y Jacquenet, F. 2015. "A first-order-logic based model for grounded language learning". En E. Fromont, T. De Bie, M. van Leeuwen (eds.), *Advances in Intelligent Data Analysis XIV*. Berlin: Springer, 49-60.

Blache, Ph. 2011. "A computational model for linguistic complexity". En G. Bel-Enguix, V. Dahl y M.D. Jiménez-López (eds.), *Biology, Computation and Linguistics. New Interdisciplinary Paradigms*. Amsterdam: IOS Press, 155-167.

Coloma, G. 2017. *La Complejidad de los Idiomas*. Bern: Peter Lang.

Dahl, Ö. 2004. *The Growth and Maintenance of Linguistic Complexity*. Amsterdam: John Benjamins.

De la Higuera, C. 2010. *Grammatical Inference: Learning Automata and Grammars*. Cambridge: Cambridge University Press.

Di Domenico, E. 2017. *Syntactic Complexity from a Language Acquisition Perspective*. Newcastle upon Tyne: Cambridge Scholars Publishing.

Givon, T. 2009. *The Genesis of Syntactic Complexity. Diachrony, Ontogeny, Neuro-cognition, Evolution*. Amsterdam: John Benjamins.

Gold, E. 1967. "Language identification in the limit", *Information and Control*, 10: 447-474.

Juola, P. 2008. "Assessing linguistic complexity". En M. Miestamo, K. Sinnemäki y F. Karlsson (eds.), *Language Complexity: Typology, Contact, Change*. Amsterdam: John Benjamins, 89-108.

Kortmann, B. y Szmrecsanyi, B. 2012. *Linguistic Complexity: Second Language Acquisition, Indigenization, Contact*. Berlin: Mouton de Gruyter.

Kusters, W. 2003. *Linguistic Complexity: The Influence of Social Change on Verbal Inflection*. Utrecht: LOT.

La Mantia, F. Licata, I. y Perconti, P. (eds.) 2017. *Language in Complexity. The Emerging Meaning*. Berlin: Springer.

McWhorter, J. 2001. "The world's simplest grammars are creole grammars". *Linguistic Typology*, 6: 125-166.

McWhorter, J. 2012. *Linguistic Simplicity and Complexity: Why do Languages Undress?* Berlin: Mouton de Gruyter.

Miestamo, M. 2008. "Grammatical complexity in a cross-linguistic perspective". En M. Miestamo, K. Sinnemäki y F. Karlsson (eds.), *Language Complexity: Typology, Contact, Change*. Amsterdam: John Benjamins, 23-42.

Miestamo, M., Sinnemäki, K. y Karlsson, F. 2008. *Language Complexity: Typology, Contact, Change*. Amsterdam: John Benjamins.

Newmeyer, F.J. y Preston, L.B. (eds.) 2014. *Measuring Grammatical Complexity*. Oxford: Oxford Univesity Press.

Pallotti, G. 2015. "A simple view of linguistic complexity", *Second Language Research*, 31: 117-134.

Pearl, L. 2010. "Using computational modeling in language acquisition research". En E. Blom y S. Unsworth (eds.), *Experimental Methods in Language Acquisition Research*. Amsterdam: John Benjamins, 163-184.

Sampson, G., Gil, D. y Trudgill, P. 2009. *Language Complexity as an Evolving Variable*. Oxford: Oxford University Press.

Trudgill, P. 2011. *Sociolinguistic Typology. Social Determinants of Linguistic Complexity*. Oxford: Oxford University Press.