UN MODELO DE APRENDIZAJE AUTOMÁTICO PARA MEDIR LA COMPLEJIDAD DE LAS LENGUAS

LEONOR BECERRA-BONACHE
Université Jean Monnet, Saint Etienne

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

RESUMEN

La invariabilidad en la complejidad de las lenguas constituye uno de los supuestos básicos de la lingüística del siglo XX. Este axioma, considerado como indiscutible durante mucho tiempo, ha sido sometido en los últimos años a un análisis exhaustivo en una serie de trabajos que defienden la conveniencia de hablar de distintos grados de complejidad lingüística. En este artículo, revisamos el concepto de complejidad y proponemos un modelo de aprendizaje automático para medir la complejidad relativa de las lenguas.

Palabras clave: complejidad lingüística, inferencia gramatical, aprendizaje automático.

ABSTRACT

The invariance of natural languages complexity is considered one of the basic assumptions of the 20th century Linguistics. Recently, this axiom, considered during many decades as an unquestioned truism in Linguistics, has been challenged in a series of works that claim that natural languages vary in complexity. In this article, we review the concept of complexity and propose a machine learning model to measure the relative complexity of languages.

Keywords: linguistic complexity, grammatical inference, machine learning.

1. INTRODUCCIÓN

A la pregunta de si todas las lenguas son igual de complejas, la lingüística del siglo XX ha respondido con el principio de la invariabilidad en el nivel de complejidad, defendiendo la hipótesis del equilibrio, que afirma que la complejidad total de una lengua es invariante porque las sub-complejidades en sub-sistemas lingüísticos se compensan. Esta idea de la *equi-complejidad,* vista durante décadas como un axioma indiscutible de la lingüística, ha empezado a ser cuestionada explícitamente en los últimos años.

Son muchos los modelos que se han propuesto para confirmar o rebatir la hipótesis de la equi-complejidad. Las herramientas, criterios o medidas que se han presentado para cuantificar el nivel de complejidad de las lenguas son muy variados y dependen de los intereses concretos de la investigación que se realice y de la definición de complejidad que se adopte. De momento, no hay una solución clara para cuantificar la complejidad lingüística y cada uno de los modelos propuestos presenta ventajas e inconvenientes.

En este trabajo, presentamos un modelo computacional que puede ayudar a calcular la complejidad lingüística. El modelo está inspirado en el proceso de adquisición del lenguaje y ha sido definido en el ámbito de la *inferencia gramatical,* subdisciplina del *aprendizaje automático.*

2. COMPLEJIDAD LINGÜÍSTICA

¿Tiene sentido comparar la complejidad de las lenguas? ¿Pueden las lenguas diferir en complejidad? Si analizamos la respuesta a estas preguntas encontramos dos tipos claramente opuestos: por un lado, quienes afirman que todas las lenguas son iguales en lo que a complejidad se refiere y, por otro, quienes consideran pertinente hablar de distintos niveles de complejidad lingüística.

El primer tipo de respuesta predomina en la lingüística del siglo XX. Durante mucho tiempo se ha defendido que la complejidad lingüística es invariante o que las lenguas son inconmensurables en lo que a complejidad se refiere y que no tiene sentido intentar demostrar que hay lenguas más complejas que otras. Estas ideas conforman lo

que algunos han denominado "dogma de la equicomplejidad lingüística" (Kusters 2003).

Ante estas ideas, las preguntas que surgen son evidentes: si las lenguas difieren en la complejidad de subsistemas particulares –cosa que admite el dogma de la equicomplejidad— ¿por qué la complejidad total es siempre la misma? ¿Qué mecanismo frena la complejidad en un área cuando ha aumentado la complejidad en otra? ¿Cuál es el factor responsable de la equicomplejidad? Estos interrogantes han hecho que recientemente se haya puesto en duda el dogma de la equicomplejidad defendiendo la posibilidad de que las lenguas presenten distintos niveles de complejidad.

Con el artículo de McWhorter (2001) en el volumen especial de la revista *Linguistic Typology* se retoma el interés por el estudio de la complejidad lingüística. A partir de este trabajo, se suceden seminarios, congresos, artículos, monografías (Dahl 2004, Kusters 2003) y volúmenes colectivos (Miestamo, Sinnemäki, Karlsson 2008; Sampson, Gil, Trudgill 2009) en los que se aborda el tema de la complejidad lingüística y se discute el dogma de la equicomplejidad, reabriendo un ámbito de estudio que ha despertado en los últimos años el interés de especialistas pertenecientes a áreas diversas.

2.1 Complejidad lingüística: Tipología

Si el objetivo es medir la complejidad de las lenguas, en primer lugar debemos determinar qué se entiende por complejidad. Este concepto ha sido interpretado de maneras diferentes en los estudios lingüísticos dedicados a este problema.

Se distingue la complejidad *absoluta* de la complejidad *relativa* (Miestamo 2008):

1. La *complejidad absoluta* se define como una propiedad objetiva del sistema. Se calcula en términos de número de partes del sistema, de interrelaciones entre las partes o de longitud de la descripción del fenómeno. Es habitual en los estudios de tipología (McWhorter 2001, Dahl 2004).
2. La *complejidad relativa* tiene en cuenta a los usuarios del lenguaje. Se identifica con la dificultad/coste de

procesamiento, aprendizaje o adquisición. Es habitual en los estudios de sociolingüística y psicolingüística (Kusters 2003).

Otra dicotomía habitual en la bibliografía es la que distingue la complejidad *global* de la complejidad *local* (Miestamo 2008):

1. La *complejidad global* calcula la complejidad total del sistema lingüístico.
2. La *complejidad local* analiza la complejidad de subdominios particulares de la lengua.

Por último, se distingue la complejidad del *sistema* de la complejidad *estructural* (Dahl 2004):

1. La *complejidad del sistema* considera las propiedades de la lengua y calcula el contenido de la competencia del hablante.
2. La *complejidad estructural* calcula la cantidad de estructura de un objeto lingüístico, analiza la estructura de las expresiones.

2.2 ¿Cómo se mide la complejidad lingüística?

En general, los estudios en este ámbito proponen medidas de complejidad *ad hoc* que dependen de los intereses del análisis realizado. Las medidas propuestas son muy variadas y pueden agruparse en dos bloques: *medidas de complejidad absoluta* como el número de categorías o reglas, longitud de la descripción, ambigüedad, redundancia, etc. (Miestamo 2008); y *medidas de complejidad relativa,* que se enfrentan al problema de determinar qué tipo de tarea (aprendizaje, adquisición, procesamiento) y qué tipo de agente (hablante, oyente, niño, adulto) considerar. La complejidad de aprendizaje de L2 (Kusters 2003) o la complejidad de procesamiento (Hawkins 2009) son ejemplos de medidas propuestas en este ámbito.

La magnitud del problema exige una solución interdisciplinar. Por ello, algunos lingüistas recurren a otras disciplinas en busca de

herramientas para calcular la complejidad lingüística. La teoría de la información (Dahl 2004, Juola 2008, Miestamo 2008), los modelos computacionales (Blache 2011), o la teoría de sistemas complejos (Andrason 2014) son ejemplos de áreas que proporcionan medidas para una evaluación cuantitativa de la complejidad lingüística.

3. COMPLEJIDAD E INFERENCIA GRAMATICAL

La variedad de propuestas muestra que no existe una solución unánimemente aceptada para cuantificar la complejidad lingüística. Considerando la necesidad de interdisciplinariedad en este ámbito, proponemos un modelo de aprendizaje automático, definido en el campo de la inferencia gramatical, para medir la complejidad relativa.

El *aprendizaje automático* se centra en el desarrollo de técnicas que permitan a los ordenadores *aprender*. Dentro de este ámbito, la *inferencia gramatical* (IG) se ocupa del aprendizaje de gramáticas/lenguajes a partir de datos. Los estudios de IG aparecen a finales de los 60 con el objetivo de formalizar el proceso de adquisición del lenguaje para intentar que una máquina consiga tal habilidad (Gold 1967). En todo problema de IG, tenemos un profesor que proporciona datos sobre el lenguaje que se quiere aprender y un aprendiz (algoritmo de aprendizaje) que debe identificar el lenguaje subyacente a partir de los datos que recibe.

Se distinguen distintos modelos de aprendizaje dependiendo de los datos que se proporcionen al aprendiz y de los criterios para determinar si ha aprendido correctamente. Los modelos más estudiados son *identificación en el límite* (Gold 1967), *aprendizaje activo* (Angluin 1987) y *aprendizaje PAC* (Valiant 1984).

En este trabajo utilizamos un modelo semántico de aprendizaje automático propuesto por Angluin y Becerra (2010).

3.1 Modelo semántico de aprendizaje de lenguas

Los estudios de IG reducen el problema de aprendizaje a la adquisición de la sintaxis, y prescinden de información semántica durante este proceso. No obstante, los lingüistas reconocen que la

información sintáctica no es suficiente para aprender una lengua y que aspectos como la semántica deben considerarse.

Angluin y Becerra (2010) proponen un modelo de inferencia gramatical, inspirado en la adquisición del lenguaje, que investiga el papel de la semántica en el proceso de aprendizaje de la lengua.

En este modelo, profesor y aprendiz interactúan en una serie de situaciones produciendo frases para denotar un objeto. Estas interacciones se desarrollan de la siguiente manera:

1. Se genera aleatoriamente una situación que se presenta al profesor y al aprendiz.
2. El aprendiz intenta producir una frase que designe uno de los objetos de esa situación.
3. El profesor produce una frase aleatoria que designa uno de los objetos de esa situación.
4. El aprendiz analiza la frase del profesor y actualiza su gramática.

Dada cualquier situación, el objetivo del aprendiz es producir frases correctas que designen un objeto en dicha situación.

Para evaluar el modelo se utilizó una simplificación de la tarea de Feldman (Stolcke 1994) conocida como *Miniature-Language-Acquisition-Task.* Esta tarea consiste en aprender un sublenguaje a partir de pares de frases-dibujos de figuras geométricas. Las lenguas consideradas fueron: inglés, alemán, griego, hebreo, húngaro, mandarín, ruso, español, sueco, turco.

En los experimentos realizados, cada situación consta de *dos objetos* y una *relación binaria* entre ambos (arriba/abajo, a la izquierda/a la derecha). Cada objeto se presenta con tres atributos: forma, color y tamaño. Se consideran 108 objetos y 23328 situaciones. El número total de significados es de 113064.

Las situaciones se describen mediante expresiones lógicas en las que se especifican las propiedades de cada objeto y la relación entre ellos. Para evaluar la actuación del aprendiz se utilizan dos medidas: 1) *exactitud*: suma de las probabilidades de las frases que se encuentran en el conjunto de frases que denotan correctamente un objeto; 2) *completitud*: fracción de las frases que denotan

correctamente un objeto que están en el conjunto de frases posibles del aprendiz. Un aprendiz alcanza un nivel *p* de actuación si exactitud y completitud son como mínimo *p*. En los experimentos, profesor y aprendiz interactúan hasta que el aprendiz consigue un nivel *p=0.99*.

La figura 1 muestra el número de interacciones necesarias para alcanzar ese nivel de actuación en las lenguas consideradas. Cada entrada es la mediana de 10 experimentos. El aprendiz es evaluado cada 100 frases recibidas del profesor.

Level	0.60	0.70	0.80	0.90	0.95	0.99
English	200	200	300	400	500	700
German	200	300	300	400	550	800
Greek	400	500	700	1500	2200	3400
Hebrew	200	300	400	500	650	900
Hungarian	200	300	350	450	550	750
Mandarin	200	200	300	400	500	700
Russian	450	500	850	1750	2350	3700
Spanish	200	300	350	500	600	1000
Swedish	200	300	300	400	600	1000
Turkish	200	200	300	400	550	800

Figura 1. Número de interacciones necesarias para que el aprendiz consiga un nivel de actuación p = 0.99. Extraída de Angluin y Becerra (2010).

En estos resultados vemos dos grupos de lenguas claramente diferenciados: por un lado, griego y ruso que necesitan al menos 3400 interacciones con el profesor; por otro, el resto de lenguas que necesitan como máximo 1000 interacciones.

3.2 Inferencia gramatical para medir la complejidad lingüística

El modelo de Angluin y Becerra parte de un algoritmo único para aprender cualquiera de las lenguas analizadas. El sistema calcula el número de interacciones necesarias para lograr un buen nivel de actuación en la lengua elegida y demuestra que no todas las lenguas necesitan el mismo número de intercambios lingüísticos para obtener el mismo nivel de adecuación.

Las características del modelo hacen que sea potencialmente adecuado para medir la complejidad *relativa* de las lenguas. Contar las interacciones necesarias para que la máquina llegue a un buen nivel de

actuación en un dominio concreto de la lengua puede verse como equivalente a calcular el coste/dificultad en el proceso de adquisición de una lengua por parte del niño. El algoritmo único utilizado en este modelo podría equivaler a la capacidad innata que capacita a los humanos para adquirir una lengua natural. Lo que demuestra el modelo es que con el mismo algoritmo no todas las lenguas requieren el mismo número de interacciones. Esto equivaldría –en términos de complejidad— a demostrar que, con la misma capacidad innata, la dificultad/coste para adquirir las diferentes lenguas no es idéntica y que, por tanto, las lenguas difieren en complejidad relativa.

El modelo, tal y como está diseñado, podría dar cuenta de la complejidad local (no global) y de la complejidad en términos estructurales (no de sistema).

4. CONCLUSIONES

Hemos presentado una primera aproximación al estudio de la complejidad lingüística con herramientas procedentes de la inferencia gramatical. Se trata de una solución interdisciplinar que recurre a un modelo computacional que permite cuantificar el coste/dificultad en el proceso de adquisición de distintas lenguas, mostrando que no es idéntico en todas ellas. Las ventajas del modelo son: su *interdisciplinariedad*, combina ideas procedentes de la lingüística con modelos computacionales; su *motivación,* es un modelo computacional basado en cómo los humanos adquieren el lenguaje; sus *resultados*, ofrece resultados experimentales cuantificables; y su capacidad para realizar *análisis croslingüísticos*.

Es necesario que la lingüística se vuelva a plantear el problema de la complejidad de las lenguas y proponga herramientas para su análisis, ya que los resultados de este tipo de estudios tendrán implicaciones importantes tanto desde el punto de vista teórico como desde el punto de vista práctico. En el plano teórico, se trata de probar o rebatir uno de los axiomas básicos de la lingüística. Desde el punto de vista práctico, el análisis de los niveles de complejidad puede ser de gran ayuda en ámbitos como la enseñanza/aprendizaje de segundas lenguas o el procesamiento automático del lenguaje.

REFERENCIAS BIBLIOGRÁFICAS

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

Angluin, D. 1987. "Learning regular sets from queries and counterexamples", *Information and Computation*, 75: 87-106.

Angluin, D., Becerra-Bonache, L. 2010. *A model of semantics and corrections in language learning.* Technical Report, Yale University.

Blache, P. 2011. "A computational model for linguistic complexity". Bel-Enguix, G., Dahl, V., Jiménez-López, M.D. (Eds.), *Biology, computation and linguistics. New Interdisciplinary Paradigms*. Amsterdam: IOS Press, 155-167.

Dahl, Ö. 2004. *The growth and maintenance of linguistic complexity.* Amsterdam: John Benjamins.

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

Hawkins, J. 2009. "An efficiency theory of complexity and related phenomena. Sampson, G., Gil, D., Trudgill, P. (Eds.), *Language complexity as an evolving variable*. Oxford: Oxford University Press, 252–268.

Juola, P. 2008. "Assessing linguistic complexity". Miestamo, M., Sinnemäki, K., Karlsson, F. (Eds.), *Language complexity: Typology, contact, change*. Amsterdam: Benjamins, 89-108.

Kusters, W. 2003. *Linguistic complexity: The influence of social change on verbal inflection.* Utrecht: LOT.

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

Miestamo, M. 2008. "Grammatical complexity in a cross-linguistic perspective". Miestamo, M., Sinnemäki, K., Karlsson, F. (Eds.), *Language complexity: Typology, contact, change.* Amsterdam: Benjamins, 23-42.

Miestamo, M., Sinnemäki, K., Karlsson, F. 2008. *Language complexity: typology, contact, change.* Amsterdam: Benjamins.

Sampson, G., Gil, D., Trudgill, P. 2009. *Language complexity as an evolving variable.* Oxford: Oxford University Press.

Stolcke, A., Feldman, J.A., Lakoff, G., Weber, S. 1994. "Miniature language acquisition: A touchstone for cognitive science", *CogSci*, 686–693.

Valiant, L.G. 1984. "A theory of the learnable", *Communication ACM*, 27: 1134-1142.