Sobre el concepto de complejidad en Lingüística

M. Dolores Jiménez López[1]
Universitat Rovira i Virgili
mariadolores.jimenez@urv.cat

Adrià Torrens Urrutia
Universitat Rovira i Virgili
adria.torrens@estudiants.urv.cat

Resumen

En este trabajo abordamos el concepto de complejidad en lingüística con un triple objetivo: revisar los usos del término y las diferentes métricas propuestas; poner de manifiesto las consecuencias teóricas y prácticas de los estudios sobre complejidad; y presentar un modelo formal para calcular la complejidad de las lenguas basado en técnicas de aprendizaje automático. La mayor parte de los estudios sobre complejidad lingüística adoptan una perspectiva absoluta y los pocos que se ocupan de complejidad relativa optan por el aprendizaje de segundas lenguas. Dada su relevancia, creemos que sería conveniente que los estudios sobre complejidad relativa consideraran el proceso de adquisición del lenguaje para determinar las diferencias entre las lenguas. Por ello, en este trabajo presentamos un modelo –que estamos desarrollando en un proyecto de investigación— basado en técnicas de aprendizaje automático que, a través de la simulación del proceso de adquisición del lenguaje, puede calcular la complejidad relativa de las lenguas.

Palabras clave: complejidad lingüística, aprendizaje automático, adquisición del lenguaje

1. Introducción

Los estudios sobre complejidad ocupan actualmente un lugar destacado en las investigaciones sobre el lenguaje natural. Este auge contrasta con la falta de atención que el concepto recibió en la lingüística del siglo XX. Durante mucho tiempo, se consideró que los estudios sobre complejidad lingüística eran innecesarios o bien porque las lenguas no diferían en complejidad, o bien porque la medición de la complejidad se consideraba irrelevante para el conocimiento de las lenguas y su funcionamiento. La falta de análisis sistemáticos para demostrar estas afirmaciones ha llevado a los investigadores a cuestionar el "dogma de la equi-complejidad" y ha dado lugar a la aparición, en los últimos años, de un gran número de trabajos que desde una perspectiva u otra intentan determinar los niveles de complejidad de las lenguas.

Un problema en los estudios sobre complejidad lingüística es la falta de acuerdo a la hora de definir este concepto. Pallotti (2015) clasifica los diferentes significados del término en tres bloques: complejidad estructural; complejidad cognitiva; complejidad del desarrollo. Estos tres significados se incluyen en los dos tipos principales de complejidad que se encuentran en la bibliografía: la complejidad absoluta, entendida como propiedad objetiva del sistema (Dahl, 2004) y la complejidad relativa, identificada con la dificultad de procesamiento, aprendizaje o adquisición (Kusters, 2003). Las diferencias en la definición comportan diferencias en las medidas propuestas y en las herramientas utilizadas para realizar el cálculo.

Los estudios sobre complejidad lingüística pueden tener consecuencias importantes. Dejar de considerar las lenguas como equicomplejas puede obligar a replantearse algunas de las cuestiones defendidas en lingüística. La evidencia de las diferencias en los niveles de

[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.

complejidad podría tener repercusiones en ámbitos como la enseñanza de segundas lenguas o las tecnologías del lenguaje.

2. Usos del término "complejidad" en lingüística

El concepto de complejidad ha sido interpretado de maneras diferentes en los estudios lingüísticos dedicados a este problema. Se distinguen distintos tipos de complejidad dependiendo del tipo de análisis que se quiera realizar.

Una de las dicotomías más repetidas es la que distingue la complejidad absoluta de la complejidad relativa (Miestamo 2008):

- La *complejidad absoluta* se define como una propiedad objetiva del sistema y se calcula en términos de número de partes del sistema, número de interrelaciones entre las partes o longitud de la descripción de un fenómeno. Esta aproximación es habitual en los estudios de tipología y está representada por trabajos como McWhorter (2001) y Dahl (2004).
- La *complejidad relativa* tiene en cuenta a los usuarios del lenguaje. Se identifica con la dificultad o coste de procesamiento, aprendizaje o adquisición. Es habitual en los estudios de sociolingüística y psicolingüística y está representada por trabajos como el de Kusters (2003).

Otra de las dicotomías habituales en la bibliografía es la que distingue la complejidad global de la complejidad local (Miestamo 2008; Edmonds 1999).

- La *complejidad global* intenta calcular la complejidad total del sistema lingüístico. Se trata de una tarea difícil que se enfrenta al problema de la representatividad –no es posible considerar de forma exhaustiva todos los aspectos relevantes de la gramática de una lengua- y al problema de la comparabilidad –la contribución de los diferentes dominios gramaticales a la complejidad global es inconmensurable.
- La *complejidad local* analiza la complejidad de subdominios particulares de la lengua y se presenta como una tarea abordable.

Por último, los estudios de complejidad lingüística distinguen la complejidad del sistema de la complejidad estructural (Dahl 2004).

- La *complejidad del sistema* hace referencia a las propiedades de una lengua, mide el número de distinciones dentro de una categoría y calcula el contenido de la competencia del hablante.
- La *complejidad estructural* calcula la cantidad de estructura de un objeto lingüístico, analiza la estructura de las expresiones.

En un artículo reciente, Pallotti (2015) subraya la polisemia del término en la bibliografía lingüística y clasifica los diferentes significados de complejidad en tres grandes bloques:

- *Complejidad estructural*: propiedad formal de textos y sistemas lingüísticos que tiene que ver con el número de sus elementos y sus patrones relacionales.
- *Complejidad cognitiva*: tiene que ver con el coste de procesamiento asociado a las estructuras lingüísticas.
- *Complejidad del desarrollo*: considera el orden en que las estructuras lingüísticas emergen y se aprenden en la adquisición de segundas (y posiblemente, primeras) lenguas.

Los tres significados anteriores cubren las dos concepciones que, según Crystal (1997), tiene el concepto en lingüística donde "complexity refers to both the internal structuring of linguistic units and psychological difficulty in using or learning them".

La mayor parte de los trabajos realizados sobre complejidad de las lenguas naturales adoptan una perspectiva absoluta del concepto y son escasos los que abordan la complejidad desde el punto de vista del usuario. Esta situación puede deberse a que ocuparse de la complejidad relativa obliga a enfrentarse a una serie de preguntas de difícil respuesta (que no se plantean cuando se aborda la complejidad como propiedad del sistema): ¿Qué significa "complejo": más difícil; más costoso; más problemático?; ¿Diferentes situaciones de uso del lenguaje (producción, comprensión, adquisición de la lengua materna, aprendizaje de segundas lenguas) pueden diferir en cuanto a lo que es difícil y lo que es fácil?; ¿Cómo decidir qué tipo de uso de la lengua (y usuario) es el primario?

Por otro lado, la mayoría de los estudios que se han ocupado de la complejidad relativa han optado por el aprendizaje de segundas lenguas, aunque algunos autores ponen en duda que el aprendizaje de segundas lenguas sea el tipo de uso más relevante (Miestamo 2006).

3. Métricas propuestas para medir la complejidad de las lenguas naturales

Para medir la complejidad, los estudios en este ámbito proponen medidas de complejidad ad hoc que dependen de los intereses concretos del análisis realizado. Las medidas propuestas son muy variadas como muestra Edmonds (1999) al enumerar 48 formulaciones diferentes. Los formalismos utilizados pueden agruparse en dos grandes bloques:

- por un lado, *medidas de complejidad absoluta* como el número de categorías o reglas, la longitud de la descripción, la ambigüedad, la redundancia etc. (Miestamo 2008);
- por otro lado, *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 en adultos (Trudgill 2001, Kusters 2003) o la complejidad de procesamiento (Hawkins 2009) son ejemplos de las medidas que se han propuesto en términos de dificultad/coste.

Algunos lingüistas han recurrido a otras disciplinas en busca de herramientas que permitan calcular la complejidad de las lenguas. La teoría de la información con formalismos como la entropía de Shannon o la complejidad de Kolmogorov (Dahl 2004, Juola 2008, Bane 2008, Miestamo 2008); los modelos computacionales basados en gramáticas de restricciones (Blache 2011); o la teoría de sistemas complejos (Andrason 2014) son algunos ejemplos de áreas que han proporcionado medidas para una evaluación cuantitativa de la complejidad lingüística.

4. Un modelo formal para medir la complejidad relativa de las lenguas naturales

El proyecto que estamos desarrollando propone una aproximación al problema de la complejidad lingüística utilizando herramientas procedentes de la inferencia gramatical (de la Higuera 2010), ámbito que se incluye dentro del aprendizaje automático.

El aprendizaje automático –subárea de la inteligencia artificial— se centra en el desarrollo de técnicas que permitan a los ordenadores aprender. Dentro de este ámbito, la inferencia gramatical estudia el aprendizaje de gramáticas y lenguajes a partir de datos.

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 ciertos paralelismos con el proceso de adquisición del lenguaje natural; en lugar de un profesor y un aprendiz, hablaríamos de un adulto y un niño (el niño aprende una lengua a partir de los datos que recibe). De hecho, los estudios de inferencia gramatical surgen a finales de los 60 con el intento de E.M. Gold (1967) de formalizar la adquisición del lenguaje. Su objetivo era formalizar el proceso de adquisición del lenguaje para poder investigar cómo una máquina podría conseguir tal habilidad.

En el proyecto que estamos desarrollando, defendemos que la inferencia gramatical puede proporcionar una buena herramienta para medir la complejidad lingüística. Los modelos propuestos en este ámbito parten de un algoritmo único para aprender cualquier lengua. El sistema calcula el número de interacciones necesarias para lograr un buen nivel de actuación en la lengua elegida y puede demostrar que no todas las lenguas necesitan el mismo número de intercambios lingüísticos para obtener el mismo nivel de adecuación.

Las características de los algoritmos de inferencia gramatical hacen que estos sean potencialmente adecuados para medir la complejidad relativa de las lenguas, esto es la complejidad entendida en términos de coste y/o dificultad de procesamiento, aprendizaje o adquisición:

- 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 o la dificultad en el proceso de adquisición de una lengua natural por parte del niño.
- El algoritmo único utilizado en estos modelos puede equivaler a la capacidad innata presente en todos los humanos que los capacita para adquirir una lengua natural.

Lo que vendría a demostrar el modelo de inferencia gramatical es que con el mismo algoritmo no todas las lenguas requieren el mismo número de interacciones. Esto equivaldría –en términos de complejidad lingüística- a demostrar que, con la misma capacidad innata, la dificultad o coste para adquirir las diferentes lenguas naturales no es idéntica y que, por tanto, las lenguas naturales difieren en complejidad relativa.

Por todo lo dicho, defendemos que los algoritmos de inferencia gramatical –entendidos como modelos computacionales de la adquisición del lenguaje— pueden ser una buena herramienta para poder considerar a los niños (y al proceso de adquisición) como usuarios (uso) adecuados para la evaluación de la complejidad de las lenguas.

5. Consecuencias teórico-prácticas de los estudios sobre complejidad

Demostrar que las lenguas difieren en complejidad puede tener consecuencias importantes tanto a nivel teórico como práctico.

En el ámbito de la lingüística teórica, se trata de rebatir uno de los axiomas básicos de la disciplina, considerado durante mucho tiempo como incuestionable. Dejar de considerar las lenguas como equicomplejas puede obligar a replantearse algunas de las cuestiones defendidas en lingüística. Áreas como la tipología lingüística, la lingüística comparativa, la lingüística histórica, la adquisición del lenguaje, etc. podrían revisar sus explicaciones a la luz de los nuevos datos, con la consecuente reformulación de los modelos y teorías tradicionalmente aceptadas. Replantearse las clasificaciones lingüísticas establecidas, comparar las lenguas partiendo de la idea de que no son iguales en lo que a complejidad se refiere, preguntarse si las lenguas en su evolución tienden a la simplificación o al aumento de

la complejidad, revisar las fases por las que el niño pasa en el proceso de adquisición del lenguaje dependiendo de la lengua que adquiera son algunas cuestiones que podrían verse afectadas por los resultados obtenidos en los estudios sobre complejidad.

En el ámbito de la lingüística aplicada, la evidencia de las diferencias en los niveles de complejidad de las lenguas naturales puede tener consecuencias interesantes en ámbitos como la enseñanza de segundas lenguas o las tecnologías del lenguaje. Quienes se dedican a la enseñanza de segundas lenguas (L2) se han interesado siempre por el concepto de complejidad como una manera de calcular el nivel de adquisición –los progresos— en el estudiante de L2. Conocer las diferencias de complejidad de las lenguas puede ayudar a plantear métodos de enseñanza/aprendizaje diferentes para cada lengua. Si las lenguas no son iguales, es probable que el mismo método de enseñanza no sea adecuado para todas ellas. De manera similar, en el ámbito de las tecnologías del lenguaje se tiende a proponer herramientas "universales" para el procesamiento automático del lenguaje que deberían funcionar para cualquier lengua natural, teniendo en cuenta que todas las lenguas son "iguales". Si se demuestra que las lenguas varían en complejidad, deberíamos tener en cuenta esas diferencias a la hora de proponer tecnología lingüística.

6. Conclusiones

Si observamos la situación de los estudios sobre complejidad lingüística, vemos que, aunque, en general, se reconoce que la complejidad es un concepto clave en lingüística, su estudio no ha sido abordado en profundidad hasta hace pocos años (Dahl 2004; Sampson et al. 2009).

Los defensores del dogma de la equicomplejidad lingüística –que son la mayoría de quienes se dedican y han dedicado a la lingüística en el último siglo— presentan tres tipos de objeciones a quienes pretenden analizar o determinar la complejidad de las lenguas:

- 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.

Que todas las lenguas tienen el mismo nivel de complejidad es algo difícil de afirmar si no se ha analizado esta cuestión en profundidad. Como hemos dicho, la lingüística ha tratado el dogma de la equicomplejidad como un axioma indiscutible y han sido prácticamente nulos los intentos de someter esta afirmación a una investigación sistemática. Por otra parte, si se analiza de forma detenida el dogma de la equicomplejidad son muchos los interrogantes que surgen: si las lenguas difieren en la complejidad de subsistemas particulares –cosa que admite el dogma de la equicomplejidad lingüística— ¿por qué la complejidad total es siempre la misma? ¿Qué mecanismo frena la complejidad en un área cuando ha aumentado la complejidad en otra? En definitiva, ¿cuál es el factor responsable de la equicomplejidad?

En lo referente a la segunda objeción, es evidente que resulta extremadamente complejo proporcionar herramientas que permitan medir la complejidad de las lenguas. De hecho, la variedad de propuestas que aparecen en la bibliografía muestra que no existe todavía una solución para cuantificar la complejidad lingüística unánimemente aceptada. Ahora bien, la existencia de un gran número de problemas a la hora de calcular la complejidad de las lenguas no implica que las lenguas no puedan ser comparadas en lo que a complejidad se refiere.

Por último, teniendo en cuenta que el análisis de la complejidad lingüística no se ha realizado de forma sistemática y que, de momento, no tenemos resultados claros sobre los niveles de complejidad de las lenguas naturales resulta precipitado afirmar que la medición de la complejidad es irrelevante para el conocimiento y funcionamiento de las lenguas.

Creemos, por tanto, que 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 pueden tener implicaciones importantes tanto desde el punto de vista teórico como desde el punto de vista práctico.

En este trabajo hemos propuesto una aproximación al problema de la complejidad lingüística utilizando herramientas de inferencia gramatical. Defendemos que los modelos computacionales de la adquisición –como los que proporcionan los algoritmos de aprendizaje automático— pueden ser una buena herramienta para poder evaluar la complejidad de las lenguas en términos relativos.

Referencias bibliográficas

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

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.

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.

Crystal, D. (1997). *The Cambridge encyclopedia of language*. Cambridge: Cambridge University Press.

de la Higuera, C. (2010). *Grammatical inference: Learning automata and grammars*. Cambridge: Cambridge University Press.

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

Edmonds, B. (1999). *Syntactic measures of complexity*. PhD.diss., University of Manchester.

Gold, E.M. (1967). ‘‘Language identification in the limit’’. *Information and Control*, 10, 447–474.

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

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.

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. (2006). ‘‘On the feasibility of complexity metrics’’. En K. Krista y M.M Sepper (eds.). *Finest Linguistics. Proceedings of the Annual Finish and Estonian Conference of Linguistics*. Tallinn: Tallinna Ülikooli Kirjastus, 11-26.

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.

Pallotti, G. (2015). ‘‘A simple view of linguistic complexity’’. *Second Language Research*, 31, 117-134.

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

Trudgill, P. (2001). ‘‘Contact and simplification: Historical baggage and directionality in linguistic change’’. *Linguistic Typology*, 5, 371-374.