Procesamiento del Lenguaje Natural, Revista nº 58, marzo de 2017, pp. 155-158 recibido 29-11-2016 revisado 11-01-2017 aceptado 02-02-2017

Análisis de la complejidad y simplificación automática de textos. El análisis de las estructuras complejas en euskera*

Readability Assessment and Automatic Text Simplification. The Analysis of Basque Complex Structures

Itziar Gonzalez-Dios
Grupo Ixa, Universidad del País Vasco (UPV/EHU)
Manuel Lardizabal 1, 20018 Donostia
itziar.gonzalezd@ehu.eus

Resumen: Tesis doctoral titulada "Euskarazko egitura sintaktiko konplexuen analisirako eta testuen sinplifikazio automatikorako proposamena / Readability Assessment and Automatic Text Simplification. The Analysis of Basque Complex Structures", defendida por Itziar Gonzalez Dios en la Universidad del País Vasco (UPV/EHU) y elaborada bajo la dirección de las doctoras Arantza Díaz de Ilarraza (Departamento de Lenguajes y Sistemas Informáticos) y María Jesús Aranzabe (Departamento de Lengua Vasca y Comunicación). La defensa tuvo lugar el 23 de junio de 2016 ante el tribunal formado por los doctores Kepa Sarasola (Presidente, Universidad del País Vasco (UPV/EHU)), Ricardo Etxepare (Secretario, Centre National de la Recherche Scientifique-IKER) y Giulia Venturi (Vocal, Instituto di Linguistica Computazionale Antonio Zampolli - Consiglio Nazionale delle Ricerche) y la tesis obtuvo la mención Cum Laude y Doctor Internacional.
Palabras clave: análisis de la complejidad o lecturabilidad, simplificación automática de textos, sintaxis, euskera

Abstract: Ph.D. thesis entitled "Euskarazko egitura sintaktiko konplexuen analisirako eta testuen sinplifikazio automatikorako proposamena / Readability Assessment and Automatic Text Simplification. The Analysis of Basque Complex Structures" written by Itziar Gonzalez Dios at the University of Basque Country (UPV/EHU) under the supervision of the Ph.D. Arantza Díaz de Ilarraza (Languages and Computer Systems Department) and Ph.D. María Jesús Aranzabe (Basque Language and Communication Department). The viva voce was held on the 23rd June 2016 and the members of the commission were the Ph.D. Kepa Sarasola (President, University of Basque Country (UPV/EHU)), Ph.D. Ricardo Etxepare (Secretary, Centre National de la Recherche Scientifique-IKER) and Ph.D. Giulia Venturi (Vocal, Instituto di Linguistica Computazionale Antonio Zampolli - Consiglio Nazionale delle Ricerche) and the thesis obtained the mention Cum Laude and International Doctor.
Keywords: readability assessment, automatic text simplification, syntax, Basque

1 Introducción de la tesis

Esta tesis doctoral se ha realizado en el grupo Ixa[1] de la Universidad del País Vasco (UPV/EHU) y se han tratado dos líneas de investigación: el análisis de la complejidad de textos o lecturabilidad (*readability assessment* en inglés) y la simplificación automática de textos (*automatic text simplification* en inglés). Concretamente, se ha analizado la complejidad sintáctica del euskera con el objetivo de diseñar un sistema de simplificación automática de textos.

Millones de textos se producen a diario en nuestra sociedad, pero estos textos no son accesibles para todos por diversos motivos: por ejemplo, algunas estructuras son complejas para las personas con enfermedades cognitivas o con alteraciones en el lenguaje y, también para las personas que aprenden lenguas extranjeras. A este último colectivo además

* Este tesis doctoral ha sido realizada con una beca predoctoral del Gobierno Vasco. Referencia: BF1-2011-392

[1] http://ixa.eus/Ixa

ISSN 1135-5948

se le añade como problema el desconocimiento del vocabulario. Pero la complejidad de los textos no es solo un problema que afecta al ser humano, sino que también afecta a las aplicaciones avanzadas del Procesamiento del Lenguaje Natural (PLN). Estas aplicaciones no procesan efectivamente oraciones largas y complejas y, por ello, mediante la simplificación automática se pretende mejorar su rendimiento.

Esta tesis tiene dos partes: 1) el análisis de las estructuras sintácticas complejas del euskera para realizar propuestas de simplificación desde un punto de vista lingüístico y 2) la producción de los recursos lingüísticos necesarios para implementar el análisis de la complejidad y la simplificación automática de textos de manera general desde una perspectiva computacional.

2 Estructura de la tesis

Con respecto a la primera parte, se dedica el primer capítulo a presentar la motivación, los objetivos y los objetos de estudio; en el segundo capítulo, se realiza un resumen de los sistemas de análisis de lecturabilidad y de los sistemas de simplificación automática explicando los diferentes tipos de simplificación, arquitecturas, técnicas y métodos de evaluación además de los recursos necesarios para ambas tareas. En el tercer capítulo se presenta el análisis lingüístico de las estructuras sintácticas complejas del euskera y se explican las propuestas para simplificar las oraciones coordinadas, yuxtapuestas, las compuestas subordinadas (sustantivas, adjetivas o relativas y adverbiales) y las aposiciones. Además, se estudian las estructuras parentéticas que contienen información biográfica que dan lugar al sistema *Biografix* que se explica en el sexto capítulo.

El cuarto capítulo se presenta como un nexo entre el análisis lingüístico que constituye la primera parte y el más computacional de la segunda parte y se expone cómo automatizar el análisis lingüístico. Para ello, se presentan las decisiones tomadas en referencia: i) a los tipos y niveles de simplificación; y ii) al algoritmo para la selección del tipo de simplificación atendiendo al usuario. Además se presentan las herramientas básicas de análisis lingüístico automático necesarias para llevar a cabo dicho proceso (la cadena de análisis de análisis desarrollada en el grupo Ixa y las herramientas básicas *Mugak* (Aranzabe, Díaz de Ilarraza y Gonzalez-Dios, 2013) y *Aposizioak* (Gonzalez-Dios et al., 2013)) desarrolladas en esta tesis.

La segunda parte está compuesta por los capítulos quinto y sexto. En el quinto capítulo se detalla el sistema de lecturabilidad *Erre-Xail* (Gonzalez-Dios et al., 2014) que discierne si un texto es simple o complejo teniendo en cuenta 96 ratios con información lingüística y técnicas de aprendizaje automático. Este sistema se utiliza como preproceso para saber si el texto de entrada debe ser simplificado o no. Si el texto es complejo, será simplificado por el sistema *EuTS* (*Euskarazko Testuen Sinplifikatzailea* [Simplificador de textos en euskera]) (Aranzabe, Díaz de Ilarraza y Gonzalez-Dios, 2012) cuya propuesta y módulos se presentan en el sexto capítulo. *EuTS* aplica las reglas lingüísticas para la simplificación sintáctica presentadas en el tercer capítulo. Como caso de estudio, se describe *Biografix* (Gonzalez-Dios, Aranzabe y Díaz de Ilarraza, 2014), que siguiendo las operaciones de *EuTS*, realiza la simplificación sintáctica de las estructuras parentéticas en euskera, castellano, alemán, francés, italiano, gallego y catalán.

Para evaluar nuestra propuesta de análisis de complejidad y simplificación, en el septimo capítulo se presenta el corpus de los textos simplificados en euskera: *Euskarazko Testu Sinplifikatuen Corpusa (ETSC)/ Corpus of Basque Simplified Texts (CBST)*. Este corpus contiene textos simplificados según las aproximaciones estructural e intuitiva y se ha anotado siguendo el esquema de anotación definido en esta tesis.

Finalmente, en el octavo capítulo se presentan las contribuciones de la tesis y el trabajo futuro. En los apéndices de la tesis, se recogen la lista de las estructuras adverbiales analizadas, las reglas de simplificación y la lista de las operaciones encontradas en ambas aproximaciones con el objeto de que sean obligatorias al aumentar el corpus o en futuras simplificaciones.

3 Contribuciones más relevantes

A continuación se presentan las contribuciones de la tesis según las dos líneas de investigación exploradas y los recursos generados.

3.1 Análisis de la complejidad y lecturabilidad

En lo referente al análisis de la complejidad, se han determinado las estructuras sintácticas consideradas complejas en euskera basándonos en los trabajos realizados en otras lenguas y en nuestro análisis lingüístico. Dichas estructuras son las oraciones coordinadas y yuxtapuestas, las compuestas subordinadas, las aposiciones y las estructuras parentéticas que contienen información biográfica. Para que las oraciones que contienen estas estructuras sean simplificadas, se ha determinado que deben tener una extensión mínima de dos sintagmas además del verbo.

A su vez, se ha implementado el sistema de análisis de lecturabilidad llamado *Erre-Xail* que determina si los textos son simples o complejos. Para ello, se han definido 96 ratios que se dividen en los grupos globales, lexicales, morfológicos, morfosintácticos, sintácticos y pragmáticos. En los experimentos realizados con la herramienta *Weka*[2], el mejor resultado de clasificación (93,50 % de precisión) se ha obtenido con la combinación de las características léxicales, morfológicas, morfosintácticas y sintácticas y con el clasificador SMO (máquinas de vectores de soporte).

3.2 Simplificación automática de textos

En lo referente a la simplificación automática de textos, se ha realizado el esquema general del sistema *EuTS*. Este sistema está basado en reglas lingüísticas, realiza dos tipos de simplificación a nivel sintáctico (sustitución sintáctica y la simplificación sintáctica) y adapta los textos a tres niveles diferentes (simplificación sintáctica superficial, simplificación natural y simplificación absoluta) adecuados a los niveles de conocimiento del euskera y a las necesidades de las aplicaciones del PLN.

En la sustitución sintáctica, las estructuras adverbiales de menor frecuencia se sustituyen por equivalentes de mayor frecuencia (Gonzalez-Dios, Aranzabe y Díaz de Ilarraza, 2015). De este modo, se consiguen los textos del nivel llamado simplificación sintáctica superficial, que son más accesibles pero que mantienen la estructura general. Este tipo de simplificación está completamente implementado y se ha evaluado cuantitativamente y cualitativamente. Cuantitativamente, el 79,63 % de las sustituciones realizadas han sido correctas, y de ellas en el 88,64 % de los casos las frases generadas has sido gramaticalmente correctas. Cualitativamente, el 75,00 % de las oraciones has resultado más fáciles de comprender para los usuarios de dicho nivel.

En la simplificación sintáctica se han establecido las operaciones y se ha recopilado toda la información lingüística necesaria para su implementación. Dichas operaciones son división, reconstrución, reordenación y correción. Mediante su aplicación se obtienen frases más cortas y, a su vez, la estructura sintáctica original desaparece. Según el nivel del usuario, los textos se adaptan a los niveles de simplificación natural o simplificación absoluta.

Asimismo, se ha implementado *Biografix* que prueba las operaciones y reglas definidas en nuestro estudio con las estructuras parentéticas que contienen información biográfica en 8 idiomas. De este modo, se ha comprobado que las reglas definidas para el euskera pueden ser adaptadas y reutilizadas en otras lenguas.

3.3 Recursos

Tres son los recursos más importantes creados en esta tesis: 1) el corpus de los textos simplificados en euskera y las herramientas básicas 2) *Mugak* y 3) *Aposizioak*.

El corpus de los textos simplificados recoge dos aproximaciones de simplificación de textos: la estructural y la intuitiva. Esto significa que cada frase original del corpus ha sido simplificada según ambas aproximaciones. Para simplificar los textos según la aproximación estructural, una traductora jurada ha seguido directrices de lectura fácil y para la aproximación intuitiva, una profesora de euskera se ha basado en su experiencia e intuición. Para realizar el análisis de este corpus se ha creado un esquema de anotación que se compone de las siguientes ocho macrooperaciones: eliminación, fusión, división, transformación, inserción, reordenación, ninguna operación y otras. Mediante este esquema, se han analizado las operaciones realizadas en cada una de las aproximaciones, y se ha creado una lista con las operaciones (distintas realizaciones de las macrooperaciones)

[2]Hall M., Frank E., Holes G., Pfahringer B., Reutemann P., y Witten I.H. The WEKA Data Mining Software: an Update. ACM SIGKDD Explorations Newsletter, 11(1):10-18, 2009.

más comunes como, por ejemplo, dividir oraciones coordinadas o recuperar los elementos elididos.

En cuanto a las herramientas básicas que se han implementado, cabe mencionar que *Mugak* detecta los límites de las oraciones basándose en información lingüística (Aranzabe, Díaz de Ilarraza y Gonzalez-Dios, 2013) y que *Aposizioak* (Gonzalez-Dios et al., 2013) detecta y clasifica las aposiciones y sus componentes. Estas dos herramientas son indispensables para realizar la operación de división.

Bibliografía

Aranzabe, M. J., A. Díaz de Ilarraza, y I. Gonzalez-Dios. 2012. First Approach to Automatic Text Simplification in Basque. En L. Rello y H. Saggion, editores, *Proceedings of the Natural Language Processing for Improving Textual Accessibility (NLP4ITA) workshop (LREC 2012)*, páginas 1–8.

Aranzabe, M. J., A. Díaz de Ilarraza, y I. Gonzalez-Dios. 2013. Transforming Complex Sentences using Dependency Trees for Automatic Text Simplification in Basque. *Procesamiento del Lenguaje Natural*, 50:61–68.

Gonzalez-Dios, I. 2014a. Euskarazko testuak errazten: euskal testuen sinplifikazio automatikoa [Making Basque Texts Easier: Automatic Simplification of Basque Texts]. En I. Aduriz y R. Urizar, editores, *Euskal Hizkuntzalaritzaren egungo zenbait ikerlerro. Hizkuntzalari euskaldunen I. topaketa*. Udako Euskal Unibertsitatea, páginas 135–149.

Gonzalez-Dios, I. 2014b. Simplificación automática de textos en euskera [Automatic Simplification of Basque Texts]. En L. A. Ureña López J. A. Troyano Jiménez F. J. Ortega Rodríguez, y E. Martínez Cámara, editores, *Actas de las V Jornadas TIMM, Cazalla de la Sierra, España, 12-JUN-2014, publicadas en http://ceur-ws.org*, páginas 45–50.

Gonzalez-Dios, I., M. J. Aranzabe, A. D. de Ilarraza, y A. Soraluze. 2013. Detecting Apposition for Text Simplification in Basque. *Computational Linguistics and Intelligent Text Processing*. Springer, páginas 513–524.

Gonzalez-Dios, I., M. J. Aranzabe, y A. Díaz de Ilarraza. 2013. Testuen sinplifikazio automatikoa: arloaren egungo egoera [Automatic Text Simplification: State of Art]. *Linguamática*, 5(2):43–63.

Gonzalez-Dios, I., M. J. Aranzabe, y A. Díaz de Ilarraza. 2014. Making Biographical Data in Wikipedia Readable: A Pattern-based Multilingual Approach. *Proceedings of the Workshop on Automatic Text Simplification - Methods and Applications in the Multilingual Society (ATS-MA 2014)*, páginas 11–20, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.

Gonzalez-Dios, I., M. J. Aranzabe, y A. Díaz de Ilarraza. 2015a. Perpaus adberbialen agerpena, maiztasuna eta kokapena EPEC-DEP corpusean [Presence, frequency and Position of Basque Adberbial Clauses in The BDP corpus]. Informe técnico, University of the Basque Country (UPV/EHU) UPV/EHU/LSI/TR 02-2015.

Gonzalez-Dios, I., M. J. Aranzabe, y A. Díaz de Ilarraza. 2015b. Simplifying Basque Texts: the Shallow Syntactic Substitution Simplification. *Proceedings of the 7th Language & Technology Conference.*, páginas 450–454.

Gonzalez-Dios, I., M. J. Aranzabe, A. Díaz de Ilarraza, y H. Salaberri. 2014. Simple or Complex? Assessing the Readability of Basque Texts. *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, páginas 334–344, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.