Simplificación automática de textos para la accesibilidad de colectivos con discapacidad: experiencias para el español y el inglés

Horacio Saggion, Universitat Pompeu Fabra, *horacio.saggion@upf.edu*

Montserrat Marimon, Universitat Pompeu Fabra, *montserrat.marimon@upf.edu*

Daniel Ferrés, Universitat Pompeu Fabra, *daniel.ferres@upf.edu*

Resumen

La simplificación de textos tiene como objetivo la transformación de un texto complejo en otro de menor complejidad de manera que este último sea más accesible para un colectivo de personas determinado. Esta simplificación usualmente se lleva a cabo reemplazando el vocabulario complicado por uno más accesible y transformando las oraciones largas y complejas en otras más cortas y simples. La simplificación automática de textos es una rama del procesamiento del lenguaje natural que investiga y desarrolla algoritmos para simplificar textos por ordenador. Si bien desde mediados de los años 90 comenzaron las investigaciones en simplificación automática de textos, esta se ha intensificado en años recientes. En este artículo presentaremos brevemente las técnicas de procesamiento de lenguaje natural que hemos implementado para desarrollar simplificadores de texto automáticos para el español y el inglés en los proyectos *Simplext* y *Able-to-Include*.

Palabras claves: simplificación automática de textos; procesamiento del lenguaje natural; simplificación sintáctica; simplificación léxica.

Abstract

Text simplification aims at transforming a complex text into a more readable and understandable version which conveys approximately the same content. The transformation usually implies the replacement of words which are difficult to understand by easy synonyms and the re-writing of long and complicated sentences into shorter ones. Automatic text simplification develops algorithms for the automation of this process and although the first works in this area appeared in the nineties, the research in this field has nowadays intensified. This article presents an overview of the natural language processing techniques we have used for the development of simplification technology for Spanish and English in the *Simplext* and *Able-to-Include* projects.

1. Introducción

Los textos que se publican a diario en portales de noticias pueden ser notablemente difíciles de leer y comprender por colectivos específicos de personas. Por ejemplo, los inmigrantes que llegan a un país pueden no tener conocimiento suficiente de la gramática y el léxico de la lengua del país de acogida y por lo tanto pueden tener dificultades en comprender algunos contenidos. Las personas con discapacidad intelectual o con alguna dificultad específica del lenguaje como las personas autistas (Evans, Orasan, & Dornescu, 2014), las personas afásicas (Carroll, Minnen, Canning, Devlin, & Tait, 1998), o las personas disléxicas (Rello, Baeza-Yates, Bott, & Saggion, 2013) pueden tener dificultades con algunas construcciones del lenguaje o con el vocabulario.

ISBN: 978-84-606-6434-5

Sin embargo, el acceso a la información es un derecho fundamental para todas las personas; en particular, en lo que respecta las personas con discapacidad, la *Convención sobre los derechos de las personas con discapacidad*[1] adoptado por las Naciones Unidas en 2006 garantiza el acceso a la información para este colectivo. Se hace necesario entonces que organizaciones gubernamentales, centros de salud, y otros organismos produzcan textos accesibles para estos colectivos de personas. Sería también importante que parte de las noticias que se publican diariamente pudieran hacerse más accesibles para estos colectivos dada la relevancia que tiene la lectura de noticias para el enriquecimiento intelectual y la integración en la sociedad. Existen varias iniciativas sobre cómo producir textos accesibles, por ejemplo la iniciativa "Plain Language" o "Plain English"[2] para el inglés o el "Basic English" (Ogden, 1932), una especie de inglés controlado con vocabulario reducido y gramática sencilla. Desde hace ya varios años se viene desarrollando para el inglés la *Simple English Wikipedia (SEW)*[3], una enciclopedia de acceso libre y en línea que contiene versiones accesibles de artículos de la Wikipedia en inglés, en teoría siguiendo los preceptos del "basic English". Existen varias organizaciones como por ejemplo la Asociación Lectura Fácil[4] que se dedican a la elaboración de textos que siguen las recomendaciones de la *Fácil Lectura* (Tronbacke, 1997). También existen varios portales de noticias que publican contenidos accesibles en español[5], francés[6], italiano[7], sueco[8], etc. Sin embargo, es sabido que producir textos accesibles es muy costoso dado el grado de especialización requerido por los editores de estos contenidos. Asimismo, el ritmo al cual las noticias se producen hace prácticamente inviable la producción de versiones accesibles para las mismas de forma manual. La simplificación automática de textos (Chandrasekar, Doran, & Srinivas, 1996)(A. Siddharthan, 2002), que viene siendo estudiada desde los años noventa, tiene como objetivo la automatización de esta tarea y podría ayudar a hacer más accesibles los contenidos textuales existentes y futuros. Dos aspectos son considerados en la simplificación automática de textos: por un lado la simplificación léxica en la que se reemplazan las palabras difíciles por sinónimos más comunes generalmente utilizando un diccionario de sinónimos. Por otro lado, la simplificación sintáctica por la cual las oraciones con construcciones lingüísticas complejas (con subordinaciones, coordinaciones, etc.) se transforman en oraciones más simples que no utilizan dichas construcciones. También se puede incluir en el proceso de simplificación, un proceso de resumen del contenido, en el cual las informaciones superfluas del texto fuente se eliminan dejando solamente las ideas principales del texto.

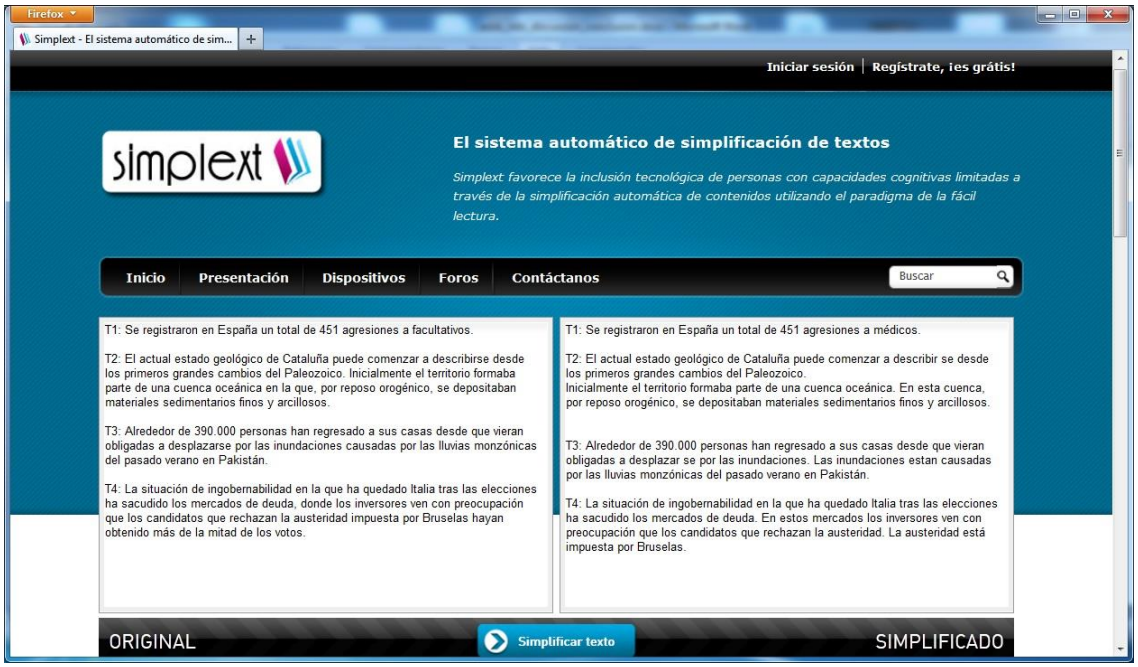
[1] http://www.un.org/spanish/disabilities/default.asp?id=497
[2] http://www.plainenglish.co.uk/
[3] http://simple.wikipedia.org/wiki/Main_Page
[4] http://www.lecturafacil.net/content-management/
[5] http://www.noticiasfacil.es/ES/Paginas/index.aspx
[6] http://cours.funoc.be/essentiel/
[7] http://www.dueparole.it
[8] http://8sidor.lattlast.se

ISBN: 978-84-606-6434-5

En este artículo presentaremos nuestro trabajo en simplificación automática de textos en dos proyectos de investigación y desarrollo que buscan avanzar el estado de la cuestión así como proporcionar herramientas viables para la inclusión de las personas con discapacidad a la sociedad de la información.

2. Proyectos *Simplext* y *Able-to-Include*

El proyecto *Simplext* (Saggion, Martínez, Etayo, Anula, & Bourg, 2011) (S. Bott & Saggion, 2014) fue el primero en desarrollar tecnología de simplificación de textos para el español. Financiado por el plan Avanza (TSI-020302-2010-84), *Simplext* se ocupó de desarrollar esta temática teniendo en cuenta un colectivo con discapacidad intelectual como las personas con síndrome Down. Uno de los desarrollos tecnológicos del proyecto es un portal Web de simplificación de textos tal como puede apreciarse en el Cuadro 1.



Cuadro 1: Portal de simplificación de textos Simplext

El proyecto *Able-to-Include*[9] tiene como objetivo el desarrollo de herramientas tecnológicas para facilitar la inclusión en la sociedad de personas con discapacidad intelectual o de desarrollo. Entre estas herramientas se destacan: (i) un sistema de simplificación de textos para el inglés y el español, (ii) un sistema de conversión de texto a pictogramas y de pictogramas a texto, y (iii) tecnología de texto a habla. Estas herramientas se harán disponibles a desarrolladores de software a través de una interfaz de programación de aplicaciones (API) para hacer posible su integración en diversas soluciones de accesibilidad. La tecnología de simplificación de textos se basa en las ideas y recursos implementados en el proyecto *Simplext.*

[9] http://abletoinclude.eu/

3 Simplificación automática en *Simplext*

La metodología adoptada en el proyecto Simplext se basó en la creación de un corpus de textos periodísticos y sus simplificaciones manuales con el fin de realizar un estudio que permitiera discernir qué manipulaciones serían necesarias para obtener una simplificación automática apropiada. Las simplificaciones manuales fueron producidas por el grupo de investigación DILES de la Universidad Autónoma de Madrid. Los diferentes estudios del corpus realizados (Stefan Bott & Saggion, 2011) (Drndarevic & Saggion, 2012) nos llevaron a proponer dos tareas principales: por un lado la *simplificación sintáctica*, encargada de transformar las frases largas y complejas en frases más simples, y por otro lado la *simplificación léxica*, encargada de simplificar el vocabulario del texto reemplazando palabras difíciles de entender por sus sinónimos más simples. También se estudiaron *operaciones de re-escritura* que hacen que los textos simplificados resulten más fáciles de entender. Cabe destacar que la creación del corpus Simplext, que será hecho público oportunamente pero que ya está disponible para los investigadores que deseen utilizarlo, es una importante contribución en el área de procesamiento de lenguaje natural en español. Algunos ejemplos de simplificaciones en el corpus Simplext pueden apreciarse a continuación.

> Original: Abre en Madrid su primera sucursal el mayor banco de China y del Mundo.
>
> Simplificación: El banco más importante de China y del mundo abre una oficina en Madrid.

Esta simplificación utiliza la palabra *oficina* en lugar de la palabra *sucursal* que es mucho más específica y por lo tanto más compleja. Además la frase simplificada utiliza el orden canónico de los elementos en la frase en español (sujeto-verbo-objeto) a diferencia del orden utilizado en la oración original.

> Original: Arranca la liga masculina de Goalball, el único deporte específico para ciegos.
>
> Simplificación: Comienza la liga masculina de Goalball. El Goalball es el único deporte específico para ciegos.

Este segundo ejemplo muestra la transformación de una aposición en la oración original en una frase independiente en el texto simplificado. Asimismo la simplificación utiliza la palabra *comienza* mucho mas común que la palabra *arranca* de la oración original.

Para desarrollar el componente de simplificación sintáctica en *Simplext* se adoptó una técnica basada en reglas que operan sobre la estructura sintáctica de dependencias de las frases (Stefan Bott, Saggion, & Mille, 2012). Las reglas son capaces de identificar y tratar construcciones de relativo, participio, gerundio, y varios casos de coordinación sintáctica. Un conjunto de reglas se encarga de identificar en las oraciones "puntos de corte" con el fin de separar las oraciones en sus componentes. Otro conjunto de reglas se encarga de copiar elementos de la frase original a los componentes más simples, por ejemplo el sujeto de la frase relativa que reemplaza el pronombre de relativo de manera que el resultado sea gramaticalmente correcto. Un tercer conjunto de reglas decide el orden en el que las diferentes frases serán presentadas. Finalmente se utiliza un sistema de generación de frases para producir las oraciones correctas.

ISBN: 978-84-606-6434-5

Con respecto al componente de simplificación léxica, se implementó el sistema LexSiS que utiliza técnicas robustas de procesamiento de lenguaje natural. LexSiS realiza una tarea de desambiguación de sentidos de las palabras en contexto para escoger sinónimos apropiados de las mismas. Por ejemplo, considere el problema de simplificar la palabra *hogar* en la oración "La madera ardía en el hogar". La palabra *hogar* tiene como sinónimos *casa* y *chimenea*, de los cuales *casa* es el más simple, sin embargo este no encaja en el contexto, de ahí que se haga necesario un proceso de desambiguación para filtrar posibles reemplazos erróneos.

Una vez desambiguada una palabra, se selecciona el sinónimo más simple para substituir la palabra original utilizando un criterio de simplicidad. Para realizar estas tareas, LexSiS cuenta con un diccionario de sentidos y sinónimos que está disponible libremente y utiliza información sobre frecuencia y longitud para seleccionar el sinónimo mas simple (Stefan Bott, Saggion, et al., 2012)(Saggion, Bott, & Rello, 2013).

El tercer componente de Simplext es un sistema de re-escritura que se encarga de transformaciones muy precisas que los simplificadores humanos aplican al texto y que no pueden ser tratadas por los componentes anteriores. Una de esas operaciones es la normalización de todos los verbos de "decir". Otras operaciones se encargan de la simplificación de expresiones numéricas tales como las fechas. Estas operaciones han sido implementadas en un sistema basado en reglas utilizando gramáticas JAPE de GATE (Maynard et al., 2002).

Los tres componentes, conjuntamente con las herramientas básicas de análisis del texto en español (etiquetadores morfosintácticos, parsers, reconocedores de entidades nombradas), han sido integrados en el motor de simplificación Simplext. Las tecnologías desarrolladas en Simplext han sido evaluadas tanto intrínsecamente como extrínsecamente con usuarios finales y no finales con métodos de evaluación propuestos en el proyecto. Detalles sobre la evaluación de Simplext pueden encontrarse en (Drndarevic, Stajner, Bott, Bautista, & Saggion, 2013).

4. Implementando un simplificador para el inglés

La simplificación de textos para el Inglés en el proyecto *Able-to-Include* aprovecha la experiencia y algunos de los recursos desarrollados en Simplext. El Simplificador Léxico para el idioma inglés es una adaptación del simplificador del español LexSiS. Este se ha desarrollado con el lenguaje de programación Java y utilizando herramientas de código abierto y recursos léxicos gratuitos para el inglés. El sistema es altamente configurable y, aunque se ha desarrollado inicialmente para personas con discapacidades intelectuales, puede ser adaptado a las necesidades específicas de diferentes grupos de usuarios solo con cambios en los recursos utilizados y los parámetros de entrada. La simplificación sintáctica en *Able-to-Include* no se basa en un estudio de corpus, sino en la identificación y tratamiento de fenómenos sintácticos que se consideran responsables potenciales de complejidad. La implementación sigue las ideas de *Simplext* en la implementación de un sistema de reglas pero utilizando formalismos diferentes.

ISBN: 978-84-606-6434-5

4.1 Simplificación léxica en *Able-to-Include*

Los componentes del simplificador léxico para el inglés son los siguientes: 1) Análisis de los documentos, 2) Detección de palabras complejas, 3) Desambiguación del sentido de las palabras, 4) Clasificación de sinónimos, y 5) Generación del lenguaje. En las siguientes secciones de desarrollan cada una de estos componentes.

4.1.1 Análisis de los documentos

Esta fase utiliza la herramienta GATE[10] y los componentes del sistema de análisis de textos ANNIE (Maynard et al., 2002) para realizar las siguientes operaciones sobre textos electrónicos: Tokenización, Segmentación de Oraciones, Etiquetado Morfosintáctico, Lematización, Reconocimiento y Clasificación de Entidades Nombradas, y Resolución de Anáfora. Estos procesos producen anotaciones lingüísticas sobre las oraciones que se utilizan en los pasos subsecuentes.

4.1.2 Detección de Palabras Complejas

La decisión sobre la complejidad de una palabra se basa en el uso de un umbral (*thresholding*) de las frecuencias de aparición de la palabra en determinados corpus o bases de datos psicolingüísticas. El procedimiento consiste en clasificar una palabra como compleja cuando la frecuencia de la palabra en la base de datos (o corpus) se encuentra por debajo de un determinado umbral. Las dos bases de datos psicolingüísticas que el sistema puede utilizar (separadamente) para la detección de palabras complejas son las *Age-of-Adquisition norms* (Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012) y los contajes de Kucera-Francis (Kučera, 1967). Las *Age-of-Acquisition norms* son puntuaciones que estiman a que edad algunas palabras representativas de la lengua inglesa (nombres, verbos, y adjetivos) son comprendidas por los humanos. La lista de frecuencias Kucera-Francis consiste en contajes de aparición de palabras en el Corpus de Brown[11] (con más de 1 millón de palabras).

Las puntuaciones de las *Age-of-Adquisition norms* se aplicaron a un conjunto total de 30.121 palabras (posteriormente expandidas a 51.715) utilizando tecnología basada en una plataforma web de colaboración abierta colectiva (*crowdsourcing*) que permitía a voluntarios escoger a qué edad (en años) creían haber comprendido las palabras. Estas puntuaciones eran originalmente de 0 a 25 (representando la edad de comprensión de la palabra), pero para nuestro sistema se invirtieron los valores de estas puntuaciones en el mismo rango de modo que, a semejanza de las frecuencias extraídas de corpus, un incremento del valor numérico normalmente representase un incremento de la simplicidad léxica.

[10] http://gate.ac.uk

[11] http://clu.uni.no/icame/brown/bcm.html

4.1.3 Desambiguación del sentido de las palabras (DSP)

En esta fase se obtiene el sentido más apropiado de una palabra compleja en un determinado contexto. El algoritmo DSP utilizado se basa en la aplicación del Modelo del Espacio Vectorial para la semántica léxica (Turney & Pantel, 2010). Esta aproximación se ha utilizado anteriormente en otros sistemas de Simplificación Léxica (Biran, Brody, & Elhadad, 2011) (Stefan Bott, Rello, Drndarevic, & Saggion, 2012).

El algoritmo de DSP empleado utiliza un modelo con vectores de palabras extraídos de colecciones de texto. Para cada una de las palabras clave que el sistema puede desambiguar, se construye un vector de palabras a partir de los contextos de aparición de la palabra en colecciones de texto. Posteriormente, se calcula un vector de palabras para cada uno de los sentidos de las palabras claves sumando los vectores de todas las palabras que pertenecen a cada sentido. Cuando una palabra compleja es detectada el algoritmo de DSP calcula la distancia (del coseno entre vectores) entre el vector de palabras del contexto de la palabra compleja (palabras de la oración o del documento completo en que se encuentra la palabra) y los vectores de palabras de cada sentido de la palabra compleja extraídos del modelo. El sentido de la palabra escogido será el que tenga la distancia del coseno mínima entre su vector de palabras en el modelo y el contexto de la palabra compleja en la oración o documento a simplificar.

Tanto las palabras clave que el sistema puede desambiguar como el diccionario para encontrar el conjunto de sentidos de las palabras se han extraído del tesauro del OpenOffice (extraído de WordNet-3.1 y originalmente de WordNet 2.0). Este diccionario contiene el conjunto de sentidos de las palabras clave según sus categorías morfosintácticas. Cada uno de los posibles sentidos de las palabras se compone de uno o más sinónimos (a veces incluyendo hiperónimos y otros términos relacionados). Este tesauro contiene 81.242 palabras clave diferentes y 135.769 entradas (palabra y categoría morfosintáctica) en el diccionario de sentidos. No se han extraído las palabras compuestas del tesauro por tanto el sistema no las simplificará.

El modelo con los vectores de palabras se ha extraído de la Simple English Wikipedia (SEW) (99.943 documentos). Este corpus se ha lematizado y etiquetado morfosintácticamente con la librería FreeLing 3.1 (Padró & Stanilovsky, 2012). El modelo de vectores de palabras se ha creado extrayendo los lemas de palabras co-ocurrentes en contextos con ventanas de tamaño 11 palabras: los 5 lemas de las palabras a la izquierda de la palabra compleja y los 5 lemas de las palabras a la derecha de la palabra compleja. Se extraen sólo las lemas de palabras de la ventana con las siguientes categorías morfosintácticas: nombres, verbos, adjetivos y adverbios.

4.1.4 Clasificación de Sinónimos

A partir de los sinónimos asociados al sentido de la palabra en el contexto (escogido en la fase anterior), esta fase clasifica estos sinónimos por su simplicidad léxica cuyo resultado es el sinónimo más simple y apropiado de la palabra compleja en su contexto original. Las medidas de simplicidad léxica implementadas son dos: a) la frecuencia de aparición de la palabra en un corpus y b) una métrica que se basa en una combinación lineal de la longitud de la palabra con la frecuencia de la palabra en un corpus (Bott et al., 2012). Para el cálculo de la medida basada en frecuencia de la palabra en corpus se

pueden utilizar varias listas de frecuencias extraídas de los siguientes corpus: British National Corpus (BNC), SEW, Wikipedia normal en inglés, Google Web 1T Corpus (Norvig's version), Age-of-Acquisition norms (invertidas), y los contajes de Kucera-Francis.

4.1.5 Generación del Lenguaje

La fase final del sistema de simplificación léxica consiste en la generación de la inflexión correcta del lema del sinónimo sustituto en el contexto de la palabra compleja. La herramienta SimpleNLG (Gatt & Reiter, 2009) se utiliza para realizar esta tarea teniendo en cuenta el lema del sinónimo sustituto, la categoría morfosintáctica de la palabra compleja original en el contexto y el contexto mismo.

4.2 Simplificación sintáctica en *Able-to-Include*

El sistema de simplificación sintáctica que hemos desarrollado simplifica las siguientes estructuras sintácticas:

(1) Tratamiento de las construcciones pasivas. Se simplifican convirtiéndolas a voz activa:

(a) *The release was accompanied by a number of TV appearances, including a full hour on "On the Record".* (El lanzamiento fue acompañado por una serie de apariciones en televisión, incluyendo una hora completa en "On the Record".)

(b) *A number of TV appearances, including a full hour on "On the Record" accompanied the release*. (Un número de apariciones en televisión, incluyendo una hora completa en "On the Record" acompañó el lanzamiento.)

(2) Tratamiento de las aposiciones. Los sintagmas nominales (SSNN) que contienen aposiciones se simplifican eliminando la aposición de la frase original y creando una nueva frase que contiene el antecedente como sujeto, el verbo ser y la aposición como predicado:

(c) *The moon is named after Portia, the heroine of William Shakespeare's play "The Merchant of Venice".* (La luna lleva el nombre de Portia, la heroína de la obra de William Shakespeare "El mercader de Venecia".)

(d) *The moon is named after Portia. Portia is the heroine of William Shakespeare's play "The Merchant of Venice".* (La luna lleva el nombre de Portia. Portia es la heroína de la obra de William Shakespeare "El mercader de Venecia".)

(3) Tratamiento de las oraciones de relativo: Los SSNN que tienen una cláusula de relativo se simplifican eliminando la cláusula de relativo de la oración original y creando una nueva cláusula en la que el antecedente sustituye al pronombre:

(e) *The festival was held in New Orleans, which was recovering from Hurricane Katrina*. (El festival se celebró en Nueva Orleans, que se estaba recuperando del huracán Katrina.)

ISBN: 978-84-606-6434-5

(f) *The festival was held in New Orleans. New Orleans was recovering from Hurricane Katrina.* (El festival se celebró en Nueva Orleans. Nueva Orleans se estaba recuperando del huracán Katrina.)

(4) Tratamiento de cláusulas y sintagmas verbales (SSVV) coordinados. Se simplifican creando una cláusula con cada uno de los elementos coordinados:

(g) *Tracy killed 71 people, caused $837 million in damage and destroyed more than 70 percent of Darwin's buildings, including 80 percent of houses.* (Tracy mató a 71 personas, causó 837 millones de dólares en daños y destruyó más del 70 por ciento de los edificios de Darwin, incluyendo el 80 por ciento de las casas.)

(h) *Tracy killed 71 people. Tracy caused $837 million in damage. And Tracy destroyed more than 70 percent of Darwin's buildings, including 80 percent of houses.* (Tracy mató a 71 personas. Tracy causó 837 millones de dólares en daños. Y Tracy destruyó más de 70 por ciento de Edificios de Darwin, incluyendo 80 por ciento de las casas.)

(5) Tratamiento de correlacionada correlativas. Se simplifican creando una cláusula con cada uno de los elementos coordinados:

(i) *A hypothesis requires more work by the researcher in order to either confirm or disprove it.* (Una hipótesis requiere más trabajo por parte del investigador con el fin de confirmarla o refutarla.)

(j) *A hypothesis requires more work by the researcher in order to confirm it. Or a hypothesis requires more work by the researcher in order to disprove it.* (Una hipótesis requiere más trabajo por parte del investigador con el fin de confirmarla. O una hipótesis requiere más trabajo por el investigador con el fin de refutarla.)

(6) Tratamiento de cláusulas adverbiales. Los SSVV modificados por cláusulas adverbiales se simplifican eliminando la cláusula adverbial y creando una nueva cláusula con el sujeto de la oración principal:

(k) *Oxfordshire is a county in the South East England region, bordering on Northamptonshire, Buckinghamshire, Berkshire, Wiltshire, Gloucestershire and Warwickshire.* (Oxfordshire es un condado en la región sudeste de Inglaterra, bordeando con Northamptonshire, Buckinghamshire, Berkshire, Wiltshire, Gloucestershire y Warwickshire.)

(l) *Oxfordshire is a county in the South East England region. Oxfordshire borders on Northamptonshire, Buckinghamshire, Berkshire, Wiltshire, Gloucestershire and Warwickshire.* (Oxfordshire es un condado en la región sudeste de Inglaterra. Oxfordshire bordea con Northamptonshire, Buckinghamshire, Berkshire, Wiltshire, Gloucestershire y Warwickshire.)

(7) Tratamiento de cláusulas subordinadas. Los SSVV modificados por cláusulas subordinadas se simplifican dividiendo la frase original en dos oraciones:

(m) *He is perhaps best known for his design for the Natural History Museum in London, although he also built a wide variety of other buildings throughout the country.* (Él es quizás más conocido por su diseño del Museo de Historia Natural

de Londres, a pesar de que también construyó una amplia variedad de otros edificios en todo el país.)

(n) *He also built a wide variety of other buildings throughout the country, but he is perhaps best known for his design for the Natural History Museum in London.* (Él también construyó una amplia variedad de otros edificios en todo el país, pero es quizás mejor conocido por su diseño del Museo de Historia Natural de Londres.)

4.2.1 Implementación

El sistema computacional está basado en reglas sintácticas de simplificación –enfoque común en los sistemas de simplificación actuales (Aluísio & Gasperin, 2010) (Advaith Siddharthan, 2006) – y simplifica las frases de forma recursiva, hasta que no se pueden aplicar más reglas de simplificación. Cuando se detectan varios fenómenos sintácticos en una misma frase, se aplica el siguiente orden de prioridad: (1) Aposición – (2) Cláusulas relativas – (3) Coordinación – (4) Correlativas – (5) Construcciones pasivas – (6) Cláusulas adverbiales – (7) Cláusulas subordinadas.

El sistema está compuesto por un proceso que incluye tres fases: (a) fase de análisis, en la que se identifica las estructuras sintácticas a simplificar, (b) fase de transformación, en la que se genera las estructuras simplificadas, y (c) fase de adaptación o de generación de lenguaje.

El análisis de los fenómenos sintácticos se realiza mediante el uso de herramientas de procesamiento del lenguaje natural específicas, que son:

- La cadena de procesamiento GATE/ANNIE, que realiza la identificación de oraciones, palabras (o expresiones multi-palabra) y entidades nombradas.
- El analizador MATE (Bohnet, 2010), que realiza el análisis de dependencias de las oraciones.
- Las reglas GATE JAPE, que detectan y analizan las diferentes estructuras sintácticas a simplificar.

La fase de transformación usa la información proporcionada por la etapa de análisis para generar frases simples.

4.3 Sistema completo *Able-to-Include*

El sistema de simplificación se ha desarrollado de manera modular lo que hace posible simplificar el vocabulario del texto, la sintaxis de las oraciones, o ambos. Asimismo, los fenómenos de simplificación sintáctica a ser tratados pueden acomodarse a las necesidades del usuario. La simplificación léxica también es flexible en cuanto a los recursos que pueden utilizarse para la determinación de la complejidad de las palabras y la selección de sinónimos.

ISBN: 978-84-606-6434-5

5 Conclusiones

La simplificación automática de textos tiene como objetivo la transformación de un texto complejo en otro de menor complejidad de manera que este último pueda ser leído y comprendido por un colectivo de personas específico. Si bien la investigación en esta área ha avanzado recientemente, los resultados obtenidos están lejos de producir simplificaciones como las de un experto humano. En este artículo hemos descrito brevemente las técnicas que hemos utilizado para la implementación de sistemas de simplificación para el español y el inglés. Nuestra tecnología es modular de manera tal que podría en principio ser utilizada para colectivos de personas con distintas necesidades. El sistema de simplificación para el español fue evaluado oportunamente identificándose las mejores que sería necesario implementar. En lo que respecta al sistema de simplificación en inglés estamos elaborando un protocolo de evaluación con usuarios, así como una comparación cuantitativa con otros sistemas existentes.

Referencias

Aluísio, R. M., & Gasperin, C. (2010). Fostering Digital Inclusion and Accessibility: The PorSimples project for Simplification of Portuguese Texts. In *Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas. : ACL* (pp. 46–53).

Biran, O., Brody, S., & Elhadad, N. (2011). Putting it Simply: a Context-Aware Approach to Lexical Simplification. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 496–501.

Bohnet, B. (2010). Very High Accuracy and Fast Dependency Parsing is Not a Contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics* (pp. 89–97). Stroudsburg, PA, USA: Association for Computational Linguistics.

Bott, S., Rello, L., Drndarevic, B., & Saggion, H. (2012). Can Spanish Be Simpler? LexSiS: Lexical Simplification for Spanish. In *Proceedings of the Conference on Computational Linguistics* (pp. 357–374).

Bott, S., & Saggion, H. (2011). Spanish Text Simplification: An Exploratory Study. *Procesamiento Del Lenguaje Natural*, *47*, 87–95.

Bott, S., & Saggion, H. (2014). Text Simplification Resources for Spanish. *Journal of Language Resources and Evaluation*.

Bott, S., Saggion, H., & Mille, S. (2012). Text Simplification Tools for Spanish (pp. 1665–1671). Proceedings of the LREC 2012 Conference.

Carroll, J., Minnen, G., Canning, Y., Devlin, S., & Tait, J. (1998). Practical Simplification of English Newspaper Text to Assist Aphasic Readers. In *In Proc. of AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology* (pp. 7–10).

Chandrasekar, R., Doran, C., & Srinivas, B. (1996). Motivations and methods for text simplification. In *PROCEEDINGS OF THE SIXTEENTH INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS (COLING '96)*.

Drndarevic, B., & Saggion, H. (2012). Reducing Text Complexity through Automatic Lexical Simplification: an Empirical Study for Spanish. *Procesamiento Del Lenguaje Natural*, *49*, 13–20.

Drndarevic, B., Stajner, S., Bott, S., Bautista, S., & Saggion, H. (2013). Automatic Text Simplification in Spanish: A Comparative Evaluation of Complementing Modules (pp. 488–500). Presented at the CICLing (2).

Evans, R., Orasan, C., & Dornescu, I. (2014). An evaluation of syntactic simplification rules for people with autism. *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, 131–140.

Gatt, A., & Reiter, E. (2009). SimpleNLG: A Realisation Engine for Practical Applications. In *Proceedings of the 12th European Workshop on Natural Language Generation* (pp. 90–93). Stroudsburg, PA, USA: Association for Computational Linguistics.

Kučera, H. (1967). *Computational analysis of present-day American English,*. Providence,: Brown University Press,.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*(4), 978–990.

Maynard, D., Tablan, V., Cunningham, H., Ursu, C., Saggion, H., Bontcheva, K., & Wilks, Y. (2002). Architectural Elements of Language Engineering Robustness. *Nat. Lang. Eng.*, *8*(3), 257–274.

Ogden, C. K. (1932). *Basic English: a general introduction with rules and grammar*. K. Paul, Trench, Trubner & Co., Ltd.

Padró, L., & Stanilovsky, E. (2012). FreeLing 3.0: Towards Wider Multilinguality. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*.

Rello, L., Baeza-Yates, R., Bott, S., & Saggion, H. (2013). Simplify or Help?: Text Simplification Strategies for People with Dyslexia. In *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility* (pp. 15:1–15:10). New York, NY, USA: ACM.

Saggion, H., Bott, S., & Rello, L. (2013). Comparing Resources for Spanish Lexical Simplification. In A.-H. Dediu, C. Martín-Vide, R. Mitkov, & B. Truthe (Eds.), *Statistical Language and Speech Processing* (pp. 236–247). Springer Berlin Heidelberg. Retrieved from

Saggion, H., Martínez, E. G., Etayo, E., Anula, A., & Bourg, L. (2011). Text simplification in simplext. making text more accessible. *Procesamiento de Lenguaje Natural*, *47*, 341–342.

Siddharthan, A. (2002). An architecture for a text simplification system. In *Language Engineering Conference, 2002. Proceedings* (pp. 64–71).

Siddharthan, A. (2006). Syntactic Simplification and Text Cohesion. *Research on Language and Computation*, *4*(1), 77–109.

Tronbacke, B. I. (1997). *Guidelines for Easy-to-read Materials*. IFLA Headquarters.

Turney, P. D., & Pantel, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics. *J. Artif. Intell. Res. (JAIR)*, *37*, 141–188.