Análisis de la Simplificación de Expresiones Numéricas en Español mediante un Estudio Empírico

Susana Bautista
Universidad Complutense de Madrid. Facultad de Informática. Madrid, España
subautis@fdi.ucm.es

Biljana Drndarević
Universitat Pompeu Fabra. Department of Information and Communication Technologies. Barcelona, España
biljana.drndarevic@upf.edu

Raquel Hervás
Universidad Complutense de Madrid. Facultad de Informática. Madrid, España
raquelhb@fdi.ucm.es

Horacio Saggion
Universitat Pompeu Fabra. Department of Information and Communication Technologies. Barcelona, España
horacio.saggion@upf.edu

Pablo Gervás
Universidad Complutense de Madrid. Facultad de Informática. Madrid, España
pgervas@sip.ucm.es

Resumen

En este artículo se presentan los resultados de un estudio empírico llevado a cabo con un corpus paralelo de textos originales y simplificados a mano, y una posterior encuesta online, con el objetivo de identificar operaciones de simplificación de expresiones numéricas en español. Consideramos una "expresión numérica" como una frase que expresa una cantidad que puede venir acompañada de un modificador numérico, como por ejemplo *casi un cuarto*. Los resultados se analizan considerando las expresiones numéricas en oraciones con y sin contexto, a partir del análisis del corpus y del análisis de los resultados recogidos en la encuesta. Consideramos como trabajo futuro llevar a cabo una implementación computacional de las reglas de transformación extraídas.

Palabras clave

Simplificación de textos, Expresiones numéricas, Estudio de corpus

Abstract

In this paper we present the results of an empirical study carried out on a parallel corpus of original and manually simplified texts in Spanish and a subsequent survey, with the aim of targeting simplification operations concerning numerical expressions. For the purpose of the study, a "numerical expression" is understood as any phrase expressing quantity possibly modified with a numerical hedge, such as *almost a quarter*. Data is analyzed both in context and in isolation, and attention is paid to the difference the target reader makes to simplification. Our future work aims at computational implementation of the transformation rules extracted so far.

Keywords

Text Simplification, Numerical Expressions, Corpus Study

1 Introducción

Debido al crecimiento de Internet, cada vez más pronunciado, existe una tendencia para digitalizar todo tipo de información con el objetivo de hacerla más accesible a los usuarios. Sin embargo, los estudios demuestran que todavía estamos lejos de ese ideal de una sociedad digitalizada uniformemente donde la información sea asequible para todos. Ciertos usuarios, como las personas con trastornos visuales o auditivos, personas con bajo nivel de alfabetización, etc., se enfrentan con dificultades a la hora de acceder al contenido digital tal y como está presentado actualmente. Por ese motivo, ha habido mucho interés últimamente, por parte de distintas instituciones internacionales, para mejorar el esta-

do de accesibilidad de contenidos que se ofrecen en la Web con el fin de incluir a grupos actualmente marginalizados. La Organización de las Naciones Unidas (ONU) postula que todo el contenido que se publica en Internet debería ser accesible para las personas con discapacidad y hace referencia a Las Pautas de Accesibilidad de Contenido Web (Web Content Accessibility Guidelines, WCAG[1]), publicadas por un grupo de trabajo de W3C (World Wide Web Consortium). Sin embargo, según un estudio llevado a cabo por la ONU[2] con el objetivo de poner a prueba el estado de accesibilidad de un conjunto de 100 páginas web del mundo, sólo tres de ellas consiguen la accesibilidad básica prescrita por WCAG.

Muchos de los contenidos en la Web se presentan en forma escrita. Por lo tanto, la estructura y nivel de complejidad del texto escrito es un factor que influye en la accesibilidad de este tipo de contenidos. Muy a menudo, textos en la Web resultan demasiado complejos e incomprensibles para ciertos grupos de lectores, entre ellos personas con discapacidades cognitivas, personas con problemas de lectura o hablantes no nativos. Ha habido varios intentos de mejorar adecuadamente el contenido de lectura, bien a través de simplificaciones de materiales ya existentes o bien escribiendo material para un grupo objetivo específico. Ése es el caso, por ejemplo, de la Simple Wikipedia en inglés (Simple English Wikipedia[3]) y la Enciclopedia Elemental Británica (Encyclopedia Britanica Kids[4]) o el portal web en español de Noticias Fácil[5]. En España, existen distintas asociaciones y programas que apoyan la promoción de la Lectura Fácil, como la Asociación Lectura Fácil [6] en Barcelona y el programa de "Vive la fácil lectura"[7] en Extremadura. La lectura fácil contempla la adaptación a un lenguaje llano de textos legales y documentos informativos para instituciones y empresas que quieran mejorar la comunicación con su público destinatario, y promueve la edición de libros para personas con dificultades lectoras. En las simplificaciones, se considera el contenido, el lenguaje, las ilustraciones, y el diseño gráfico.

Sin embargo, la simplificación manual es demasiado lenta y costosa para ser una forma efectiva de producir la suficiente cantidad de material de lectura deseado. Por esta razón ha habido numerosos intentos de desarrollar sistemas de simplificación de textos automáticos o semi-automáticos, principalmente aplicados al inglés (Medero y Ostendorf, 2011), pero también japonés (Inui et al., 2003), portugués (Specia, 2010) y ahora español (Saggion et al., 2011). Estos sistemas utilizan técnicas computacionales en conjunto con los recursos lingüísticos para tratar tanto la estuctura sintáctica como el vocabulario del texto original que se ha de simplificar.

Nuestro trabajo sigue esta línea de investigación y se centra en esta contribución en las estrategias de simplificación léxica en textos informativos de género periodístico en español, con el objetivo de hacerlos más accesibles a las personas con discapacidad cognitiva. La importancia de las operaciones léxicas en la simplificación de textos ha sido ya tratada en trabajos previos (Carroll et al., 1998), (De Belder, Deschacht, y Moens, 2010), (Specia, 2010). El análsis del corpus que hemos llevado a cabo para el próposito de este artículo muestra también que los cambios léxicos son el tipo más común de todas las operaciones que aplican los editores humanos a la hora de simplificar un texto. En términos generales, las palabras y expresiones que se perciben como complicadas se cambian por sus sinónimos más simples o se parafrasean, como en el ejemplo que sigue (1 es la frase original, y 2 su simplificación)[8]:

1. *El Consejo de Ministros* **ha concedido** *hoy la Orden de las Artes y las Letras de España al* **restaurador** *José Andrés, a la escritora* **estadounidense** *Barbara Probst Solomon y al* **psiquiatra** *Luis Rojas Marcos.*

2. *Hoy el Gobierno de España* **ha dado** *el premio de la Orden de las Artes de España a tres personas. Al* **cocinero** *José Andrés, a la escritora* **de Estados Unidos** *Barbara Probst Solomon y al* **médico** *Luis Rojas Marcos.*

El primer cambio significativo es que la frase original ha sido divida en dos frases simplificadas. Además, en negrita se muestran los cambios observados en cuatro unidades léxicas.

En este trabajo nos centramos en un tipo particular de expresiones léxicas - las que con-

[1] http://www.w3.org/TR/WCAG/ [Último acceso: 20/11/2012]

[2] http://www.un.org/esa/socdev/enable/gawanomensa.htm [Último acceso: 20/11/2012]

[3] http://simple.wikipedia.org/wiki/Main_Page [Último acceso: 20/11/2012]

[4] http://kids.britannica.com/ [Último acceso: 20/11/2012]

[5] http://www.noticiasfacil.es/ES/Paginas/index.aspx [Último acceso: 20/11/2012]

[6] http://www.lecturafacil.net/content-management-es/ [Último acceso: 20/11/2012]

[7] http://www.facillectura.es/ [Último acceso: 20/11/2012]

[8] El ejemplo está extraído del corpus que describimos en la Sección 3.1

tienen información numérica. Consideramos una "expresión numérica" (ExpNum) como una frase que expresa una cantidad, opcionalmente acompañada de un modificador numérico, como son las expresiones: *más de un cuarto* o *cerca del 97 %*, donde *más de* y *cerca de* son ejemplos de modificadores numéricos. Este tipo de expresiones aparecen con una elevada frecuencia en el tipo de textos periodísticos que tratamos. A menudo las noticias diarias contienen información en forma numérica, y el modo en el que se presenta esta información afecta a la legibilidad de dichos textos. Consideremos la siguiente noticia, parte del corpus Simplext (ver Sección 3.1), y fijémonos en el número y la variedad de expresiones numéricas que contiene (marcadas en negrita):

> **CASI 400.000 PERSONAS** *DESPLAZADAS EN PAKISTÁN HAN VUELTO A CASA TRAS LAS INUNDACIONES*
>
> **Alrededor de 390.000 personas** *han regresado a sus casas desde que se vieran obligadas a desplazarse por las inundaciones causadas por las lluvias monzónicas del pasado verano en Pakistán. Según la Oficina de la ONU para la Coordinación de Asuntos Humanitario, esta cifra supone* **un 26 %** *de los* **1,5 millones de pakistaníes** *desplazados por las inundaciones. Por otro lado, la ONU ha logrado recaudar* **un 34 %** *de los* **2.000 millones de dólares** *(***cerca de 1.400 millones de euros***) solicitados como llamamiento de urgencia ante la catástrofe de Pakistán, la mayor petición realizada nunca por Naciones Unidas ante un desastre natural. Esta catástrofe ha matado a* **unas 2.000 personas**, *ha afectado a* **más de 20 millones**, *ha destruido* **cerca de 1,9 millones de hogares** *y ha devastado* **al menos 160.000 kilómetros cuadrados**, **una quinta parte del pas**. *Ante esta tesitura, el secretario general de la ONU, Ban Ki-moon, ha urgido a la comunidad internacional a responder "con generosidad y rapidez" a las necesidades humanitarias de Pakistán.*

En un texto relativamente corto encontramos hasta 12 expresiones numéricas distintas, que suponen dos expresiones numéricas por frase, en términos medios. Tanta carga informativa, al igual que la variedad de expresiones numéricas diferentes, pueden interferir con la comprensión del texto e impedirle al lector descubrir las relaciones de causa y efecto de los acontecimientos tratados en la noticia.

Por eso decidimos centrarnos en el tratamiento de las expresiones numéricas para la simplificación de textos en español. Este es un tema que no ha sido tratado en la literatura hasta ahora. Empezamos con un análisis de corpus, en el que observamos los cambios relativos a expresiones numéricas, hechos por humanos. De dicho corpus extrajimos un conjunto de expresiones numéricas y las presentamos en una encuesta, para que un grupo de participantes las simplificaran fuera de su contexto original. Nuestro objetivo es obtener un conjunto de operaciones para la simplificación de expresiones numéricas y plantear su implementación computacional, que sería una de las tareas en el proceso de simplificación de textos.

Este artículo está organizado como sigue: la Sección 2 presenta los trabajos relacionados en este área; en la Sección 3 describimos el conjunto experimental del estudio; el análisis de los datos es descrito en la Sección 4; la Sección 5 recoge nuestra discusión y conclusiones. Las líneas de trabajo futuro son presentadas en la Sección 6.

2 Trabajo Previo

Hasta ahora la simplificación de textos ha sido enfocada con dos objetivos diferentes. Uno es ofrecer versiones simplificadas de textos originales a grupos específicos de lectores humanos, como:

- estudiantes de lenguas extranjeras (Medero y Ostendorf, 2011);
- personas afásicas (Carroll et al., 1998), (Devlin y Unthank, 2006);
- personas con discapacidad auditiva (Inui et al., 2003);
- personas con bajo nivel de alfabetización (Specia, 2010), (Candido et al., 2009);
- personas no familiarizadas con textos técnicos altamente idiosincráticos tales como las patentes y los reglamentos (Bouayad-Agha et al., 2009).

Por otro lado, la simplificación de textos podría mejorar la eficiencia de otras tareas del procesamiento del lenguaje natural, tal y como se ha visto en los sistemas de traducción automática o en los sistemas de extracción de información (Chandrasekar, Doran, y Srinivas, 1996), (Klebanov, Knight, y Marcu, 2004).

De cualquier manera, la simplificación de texto hasta ahora ha afectado principalmente a las

construcciones sintácticas y a las expresiones léxicas percibidas como complejas o complicadas, como son oraciones largas con múltiples oraciones coordinadas y subordinadas, oraciones en voz pasiva, uso de palabras de baja frecuencia, palabras abstractas, términos técnicos y abreviaturas. Chandrasekar, Doran, y Srinivas (1996) y Siddharthan (2002) se centran principalmente en estructuras sintácticas, mientras que Carroll et al. (1998), dentro de su proyecto PSET (Practical Simplificacion of English Text) orientado hacia lectores con afasia, introducen también un módulo de simplificación léxica. Su enfoque se basa en búsqueda de sinónimos en WordNet en combinación con las frecuencias Kucera-Francis, extraídas de la base de datos Oxford Psycholinguistic Database (Quinlan, 1992). Por lo tanto, el sinónimo con mayor frecuencia dentro del conjunto de sinónimos extraídos para cada palabra léxica del texto original se escoge como su equivalente más simple.

Dicho enfoque basado en sinonimia y frecuencia de palabra ha sido reutilizado en varios trabajos. Lal y Ruger (2002) utilizan el mismo método para el componente léxico de su sistema de resumen automático. Burstein et al. (2007) se centran en los cambios de vocabulario a la hora de ofrecer su sistema ATA V.1.0 como herramienta para la adaptación de textos, pensada para los profesores y estudiantes de lenguas extranjeras. Su sistema produce párrafos resumidos del texto original, llamados notas marginales, y al mismo tiempo le ofrece al usuario sinónimos más frecuentes de palabras poco usadas, extraídos de WordNet calculando la similitud de palabras. Bautista, Gervás, y Madrid (2009) también emplean diccionarios de sinónimos, pero su criterio para escoger el más adecuado es longitud de palabra, en vez de la frecuencia.

Dado que muchas palabras, en particular las palabras con mayor frecuencia, tienden a ser polisémicas, se han visto varios intentos de tratar este problema con el objetivo de conseguir una sustitución léxica más precisa que también tenga en cuenta el contexto. Con este fin, De Belder, Deschacht, y Moens (2010) fueron los primeros en utilizar técnicas de desambiguación del sentido de las palabras. Para cada palabra léxica se crean dos conjuntos de "palabras alternativas" uno basado en sinónimos de WordNet o algún diccionario parecido, y otro generado con el modelo de lenguaje del análisis semántico latente (Deschacht y Moens, 2009). Una vez determinada la intersección de estos dos conjuntos, se calcula la probabilidad para cada palabra de la intersección con el fin de comprobar si dicha palabra es un reemplazo adecuado para la palabra de entrada. La probabilidad se calcula teniendo en cuenta la dificultad de la palabra basada en la frecuencia Kucera-Francis, el número promedio de sílabas y la probabilidad de cada palabra extraída de un corpus de textos de fácil lectura, tal como la Simple English Wikipedia.

Biran, Brody, y Elhadad (2011) emplean un método no supervisado de aprendizaje automático para aprender pares de sinónimos de palabras complejas y simples, basado en un corpus no alineado de textos de la Wikipedia original y la Wikipedia simple en inglés. Yatskar et al. (2010) también utilizan un método no supervisado para extraer simplificación léxica, utilizando el historial de ediciones de la Wikipedia simple en inglés.

En cuanto a las expresiones numéricas, existen algunos trabajos, aunque dirigidos principalmente a los expertos y no a los individuos con dificultades numéricas (Peters et al., 2007), (Dieckmann, Slovic, y Peters, 2009), (Mishra H, 2011).

Bautista et al. (2011) y Power y Williams (2012) se encuentran entre los primeros en concentrarse en la posibilidad de simplificar este tipo de expresiones, centrándose principalmente en el uso de modificadores. Power y Williams (2012) realizaron un estudio de un corpus de noticias en inglés, analizaron como los autores variaban las formas matemáticas y la precisión de las mismas cuando ellos expresaban información numérica. En un documento una misma cantidad era a menudo descrita de distintas maneras, variando su expresión (fracción, porcentaje) y su precisión, usando modificadores y redondeo para ello. Además, desarrollaron un sistema basado en restricciones para decidir como adaptar la proporción original. El trabajo de Bautista et al. (2011) estudia la preferencia de valores comunes a la hora de redondear las expresiones numéricas y el uso de diferentes estrategias de simplificación dependiendo del valor de la proporción original. Está desarrollado para textos en inglés, no fue dirigido a un grupo determinado de lectores, y la simplificación se realizó de acuerdo a los niveles de dificultad según se describen en el Currículo de Matemáticas de la Autoridad de Calificaciones y Currículum de Inglaterra (Qualifications y Authority, 2010).

3 Metodología y Objetivos

Con el objetivo de esbozar conclusiones sobre el tipo de operaciones de simplificación que podrían ser aplicadas a las expresiones numéricas, hemos llevado a cabo un estudio de un corpus paralelo de textos originales en español y su

correspondiente versión simplificada a mano. El estudio del corpus forma parte de un trabajo más amplio, cuyo objetivo es desarrollar un sistema para la simplificación autómatica de noticias en español. Desarrollando el módulo de la simplificación léxica, hemos observado un número elevado de expresiones numéricas y sus simplificaciones en el corpus. En un intento de investigar más a fondo el caso de la simplificación de dichas expresiones, las tratamos como un caso específico de la simplificación léxica y las analizamos por separado.

Con el fin de ampliar el conjunto de las posibles simplificaciones relacionadas a estas expresiones, llevamos a cabo una encuesta complementaria al estudio del corpus. Las expresiones numéricas del corpus han sido etiquetadas y extraídas, junto con el resto de la frase donce aparecen, para presentarlas de manera separada en dicha encuesta. A los participantes de la encuesta se les pidió que simplificaran las expresiones numéricas que se les ofrecieron.

Por lo tanto, por un lado tenemos expresiones numéricas en contexto, es decir, en el corpus, donde se pueden observar otras operaciones de simplificación, como por ejemplo sustituciones basadas en sinonimia o reestructuración sintáctica. Además de eso, el corpus fue simplificado por expertos teniendo en mente como usuario final un lector específico - una persona con dificultades lectoras debido a discapacidades cognitivas. Por otro lado, se extrajeron oraciones individuales del mismo corpus que contienen expresiones numéricas, y se presentaron fuera de contexto a los participantes de la encuesta para que las simplificaran, sin tener en cuenta quién era el usuario final. El objetivo es ampliar el conjunto de posibles operaciones de simplificación de las expresiones numéricas, no necesariamente relacionadas a un género de texto o a un usuario final dado. En el caso de la encuesta, estas simplificaciones fueron libres, en el sentido que fueron simplificadas sin especificar ningún grupo objetivo de lectores, por lo que los participantes simplificaron de manera general.

Dentro de la variedad de tipos encontrados en las expresiones numéricas, hemos limitado nuestro trabajo al tratamiento de expresiones monetarias (*15 millones de euros*), porcentajes (*24 %*), fracciones (*un cuarto*), dimensiones físicas (*160,000 kilómetros cuadrados*) y cantidades generales (*2,000 personas*). En la sección 4.3 se discute cómo las simplificaciones hechas en el corpus y en la encuesta difieren y se complementan unas a otras, con la intención de obtener conclusiones para la posible implementación computacional de la simplificación de expresiones numéricas. A continuación describimos el conjunto de datos experimental, al igual que los recursos empleados para el análisis - el corpus, las herramientas del procesamiento del texto y la encuesta.

3.1 Corpus

Como parte de un proyecto más amplio[9], orientado hacia el desarrollo de un sistema de la simplificación automática de textos en español para los lectores con discapacidad cognitiva, hemos recopilado un corpus paralelo para usar como base para un análisis empírico. Dicho corpus consiste en 40 textos informativos, en el dominio de noticias internacionales y de cultura, cedidos por la agencia española de noticias Servimedia[10]. Los textos han sido simplificados por editores humanos, teniendo en cuenta el usuario final - un lector con discapacidad cognitiva, y siguiendo una serie de pautas de la metología de fácil lectura sugerida por Anula (2007), (2008). Dichas pautas incluyen una serie de reglas, que se podrían resumir de la siguiente manera:

- tratamiento de la microestructura del texto, es decir la estructura de la frase y los elementos del vocabulario;
- tratamiento de la información, como la reducción o expansión del contenido;
- tratamiento del discurso, como el estilo;
- la aplicación de una adecuada norma ortográfica.

Ambos conjuntos de textos, original y simplificado, han sido anotados automáticamente usando las etiquetas del procesamiento morfológico de las palabras, el reconocimiento de entidades nombradas y el análisis sintáctico, proporcionados por el paquete de análisis de lenguaje de FreeLing (Padró et al., 2010), descrito con más detalle en la sección 3.2. Además de esto, un algoritmo de alineación de textos (Bott y Saggion, 2011) ha sido aplicado para conseguir alineación a nivel de oración entre los textos originales y simplificados. Los errores de alineación han sido manualmente corregidos usando una herramienta gráfica de edición en el marco de GATE (General Architecture for Text Engineering) (Maynard et al., 2002).

[9] www.simplext.es [Último acceso: 20/11/2012]

[10] http://www.servimedia.es/ [Último acceso: 20/11/2012]

De esta manera hemos obtenido un corpus paralelo de un total de 570 oraciones, 246 en el conjunto original y 324 en el conjunto simplificado. Dicho corpus nos ha servido para documentar todas las operaciones de edición aplicadas por los humanos para planificar y organizar su implementación automática. Entre la variedad de operaciones detectadas actualmente nos centramos en simplificaciones léxicas, más específicamente en el tratamiento de las expresiones numéricas, que es el trabajo que presentamos en este artículo.

3.2 Procesamiento del texto

Tal y como mencionamos en el párrafo anterior, los textos del corpus han sido analizados usando FreeLing (Padró et al., 2010) y después procesados con la herramienta de edición de textos GATE (General Architecture for Text Engineering) (Maynard et al., 2002). GATE es un conjunto de herramientas para el procesamiento de lenguaje natural que se integran en una plataforma escrita en Java. Dispone de una interfaz gráfica y un entorno de desarrollo integrado que facilita considerablemente las tareas que requieren un proceso de edición y editores especializados. GATE es de distribución libre y de código abierto.

FreeLing es una de las herramientas de análisis del procesamiento de lenguaje natural existentes para el castellano que permite realizar análisis morfológico (part-of-speech tagging) basado en un modelo de Markov con estados ocultos. Este tipo de análisis anota los textos e identifica los lemas de cada palabra, asignándole su correspondiente etiqueta. El sistema de etiquetado usado por FreeLing sigue el estándar EAGLES[11]. Para el propósito de este artículo nos hemos centrado en las etiquetas correspondientes a expresiones numéricas. A las cifras y a los números se les asigna la etiqueta Z. Bajo esta etiqueta podemos encontrar números, ratios, porcentajes, dimensiones, etc. FreeLing identifica cuatro tipos distintos de numerales que etiqueta de manera distinta:

1. Los numerales partitivos tienen la etiqueta Zd (p.e. *una docena*, *un millón*, *un centenar*, etc.).

2. Las cantidades monetarias reciben la etiqueta Zm, que tienen como lema la cantidad (en cifras) y el nombre de la unidad monetaria en sigular (p.e. *2000 dólares*, cuyo lema es $_USD:2000)

3. Las fracciones y porcentajes tienen la etiqueta Zp. El lema normaliza la proporción (p.e. *74 %*, cuyo lema es 74/100)

4. Las magnitudes físicas reciben la etiqueta Zu. El lema normaliza la unidad de medida y la magnitud (p.e. *30Km/h*, cuyo lema es SP_km/h:30).

Para empezar, usamos FreeLing para el análisis morfológico del corpus, y una vez que los textos están etiquetados, llevamos a cabo la tarea de anotación de las expresiones numéricas en GATE. Para hacer posible la integración de ambas herramientas, es necesario convertir el formato de salida de FreeLing en un formato XML legible por GATE.

Para anotar las diferentes expresiones numéricas en los textos originales, incluyendo sus posibles modificadores, hemos utilizado GATE para definir un conjunto de gramáticas JAPE (Java Annotation Patterns Engine). JAPE es una versión de CPSL - Common Pattern Specification Language. JAPE proporciona la traducción de estados finitos sobre anotaciones basadas en expresiones regulares y reconoce las expresiones regulares en las anotaciones en los textos que queremos analizar. Una gramática JAPE contiene conjuntos de reglas, organizadas en fases y compuestas por patrones y sus correspondientes acciones. Las fases se ejecutan en cascadas de transductores de estados finitos sobre las anotaciones en los textos originales. La parte izquierda de la regla (left-hand-side, LHS) describe el patrón de la anotación, mientras la parte derecha de la regla (right-hand-side, RHS) sirve para declarar qué acciones ejecutar sobre la anotación en cuestión. Es posible hacer referencia a las anotaciones de LHS en la parte de la derecha, poniéndoles etiquetas a los elementos del patrón.

En la Figura 1 se puede ver un ejemplo de un texto original del corpus con las expresiones reconocidas usando las gramáticas JAPE definidas para anotar los distintos tipos de expresiones numéricas. El Cuadro 1 muestra un ejemplo de la regla titulada “CasiPorcFract”, que usamos para identificar las expresiones numéricas de tipo porcentajes y fracciones acompañadas por el modificador “casi”. La parte que precede a “–>” es la parte izquierda, y la parte derecha es la parte que le sigue. La parte izquierda especifica un patrón que tiene que coincidir con las anotaciones que existen en el documento GATE, mientras que la parte derecha especifica que es lo que hay que hacer con el texto coincidente. En el ejem-

[11] http://nlp.lsi.upc.edu/freeling/doc/tagsets/tagset-es [Último acceso: 20/11/2012]

Figura 1: Ejemplo de texto con las expresiones reconocidas usando gramáticas JAPE

plo, la regla tiene el título "CasiPorcFract", la cual comprueba en el texto anotado las palabras que tienen en su lema una característica "casi" y la palabra está anotada con la etiqueta "Zp". Una vez que la regla ha encontrado una secuencia de texto que coincida con este patrón, la anota con la etiqueta que se indica después de la palabra "annotate" en la parte derecha de la regla, en este caso, con la etiqueta "CASIporcFract". Además, dentro de la expresión numérica identificada, se etiqueta como MOD_EXP el texto que corresponde con el modificador y que ha sido identificado en la parte izquierda con la etiqueta "modifier". De esta forma, tendremos anotado dentro de la expresión numérica tanto el modificador como la cantidad. El texto queda anotado con la gramática JAPE definida para este tipo de expresión "CasiPorcFract", cuyo modificador es "casi", acompañado de cualquier cantidad etiquetada por "Zp", como se puede ver en el ejemplo de la Figura 2, para el caso de "Casi el 20 %".

Rule: CasiPorcFract
(((word.lemma="casi") (word)?): **modifier**
(word.tag="Zp")):**annotate**
–>
:**modifier**.MOD_EXP={semantics="casi"},
:**annotate**.CASIporcFract= {semantics="porcFract"}

Cuadro 1: Ejemplo de una regla de una gramática JAPE

Estas gramáticas en GATE las usamos para

Figura 2: Ejemplo de expresión numérica anotada correspondiente a la regla JAPE mostrada

anotar todos los distintos tipos de expresiones numéricas que encontramos en el corpus. Esto nos permite llevar a cabo un análisis del corpus e identificar diferentes tipos de expresiones para ser presentadas en la encuesta a los participantes. Para desarrollar las reglas hemos contado con el sistema ANNIC (Aswani et al., 2005), y un componente de GATE para indexación, anotación y búsqueda. Este sistema nos permite hacer búsqueda en el corpus anotado con las etiquetas de nuestro interés, que han sido generadas a partir de las reglas que hemos definido en nuestras gramáticas. Este conjunto de gramáticas JAPE es un primer paso para una futura implementación de las reglas de simplificación.

Sobre un subconjunto de 10 textos, con un total de 59 oraciones, pertenecientes al corpus se lleva a cabo la correción manual de las reglas ejecutadas automáticamente. Usando la herramienta GATE se hace una com-

paración automática identificando las etiquetas nuevas creadas manualmente y las generadas automáticamente a partir de las gramáticas JAPE definidas. Las gramáticas desarrolladas utilizando el método previamente explicado tienen una cobertura de 13 casos diferentes de expresiones numéricas de los cuatro tipos distintos identificados por el analizador. En el Cuadro 2 mostramos los 13 casos identificados en el corpus usado para medir la cobertura de las reglas definidas.

Hemos comprobado el rendimiento de las reglas definidas y hemos obtenido los siguientes resultados globales: precision= 0.94, recall= 0.93 y F-measure= 0.93. Para cada etiqueta, GATE calcula, precision, recall y F-measure, y hemos observado que en las expresiones numéricas menos frecuentes se obtienen peores resultados pero para las expresiones numéricas más frecuentes se obtienen muy buenos resultados. En los resultados globales vemos que tenemos una precision y un recall muy altos, ya que nuestras reglas etiquetan una fracción bastante alta de las instancias relevantes del corpus.

3.3 Encuesta

El objetivo de la encuesta es ampliar el conjunto de posibles operaciones de simplificación de expresiones numéricas obtenidas del corpus. Oraciones aisladas que contienen expresiones numéricas se les ofrecen a los participantes en la encuesta para que propongan sus propias simplificaciones.

Para ello, se preparó un cuestionario usando la herramienta que proporciona Google para hacer formularios, y se albergó en Google Docs[12]. La evaluación experimental incluyó a 23 participantes, todos hablantes nativos de español en posesión de un título universitario. El cuestionario se compone de frases tomadas de la recopilación antes mencionada, con la diferencia de que el contexto que las rodea fue omitido y el único cambio que se aplica es el relativo a las expresiones numéricas que se traten en cada oración. Para este cuestionario se optó por 14 frases con un total de 27 expresiones numéricas. Doce de las expresiones originales ya contenían un modificador, mientras que las 15 restantes no lo contenían. La siguiente frase es un ejemplo del tipo de oraciones que se presentaron en la encuesta:

> *Esta catástrofe ha matado a [***unas 2.000 personas***], ha afectado a [***más de 20 millones***], ha destruido [***cerca de 1,9 millones de hogares***] y ha devastado [***al menos 160.000 kilómetros cuadrados***], una [***quinta parte***] del país.*

Los participantes tenían que proporcionar simplificaciones de las expresiones numéricas marcadas por corchetes en cada frase que se presentaba en el cuestionario. Las instrucciones decían que las expresiones numéricas se podían simplificar utilizando cualquier formato: números en palabras, cifras, fracciones, proporciones, etc. Así mismo, se indicó que los modificadores tales como *menos que* o *alrededor de* podían ser utilizados si se consideraba necesario. A los participantes se les indicó que mantuvieran el sentido de la frase en la versión simplificada tan cerca como fuese posible del sentido de la oración original y que, de ser necesario, se podía reescribir la sentencia original completa. No se impusieron más restricciones, es decir, los usuarios no recibieron instrucciones para aplicar las reglas de simplificación que se habían extraído previamente del corpus, dado que la idea era compararlas con las operaciones extraídas del corpus y estudiar dicha comparación. La Figura 3 muestra una pequeña parte de la encuesta, donde se puede ver una oración que se presentó a los usuarios, con una expresión numérica entre corchetes, la cuál se pedía simplificar.

4 Análisis de los datos

Aquí presentamos los resultados obtenidos por separado: en primer lugar, a partir del análisis del corpus, y en segundo lugar, a partir del análisis de los resultados recogidos en la encuesta realizada. Los datos obtenidos se analizan con un enfoque comparativo, con el objetivo de extraer conclusiones sobre la posibilidad de la implementación de las reglas de simplificación extraídas.

4.1 Análisis del corpus

Como ya se ha mencionado, aquí tratamos expresiones numéricas como casos específicos de simplificación léxica. El análisis del corpus, compuesto por textos periodísticos, que se llevó a cabo con el fin de extraer las estrategias de simplificación léxica, ha mostrado que las expresiones numéricas no sólo son abundantes en este género, sino que también se modifican con frecuencia para conseguir un texto de salida más fácil de leer. Cada texto original contiene un promedio de 3,78 expresiones numéricas.

[12] https://docs.google.com/spreadsheet/viewform?formkey=dDhWQ2NyckpUTUthbTVIRVVFTUtaRGc6MQ#gid=0 [Último acceso: 20/11/2012]

Etiqueta	Expresión Numérica	Ejemplo
CASIporcFract	casi + Zp	casi un cuarto
DURANTENUM	durante + Z	durante 24 das
MASDENUM	más de + Z	más de 50.000
MASDEPART	más de + Zd	más de 20 millones
MASDEporcFract	más de + Zp	más del 40 %
NUMERALES	Z	34.589
NUMMAGNITUDES	Zu	32 metros
NUMMONETARIAS	Zm	1.400 euros
NUMPARTITIVO	Zd	32 millones
NUMPORCENTYFRACT	Zp	75 %
UNASMagnit	unas + Zu	unas 700 millas
UNASNUM	unas + Z	unas 20.000
MOD_EXP	modifier	alrededor, menos de...

Cuadro 2: Tipos identificados en el corpus usado para medir la cobertura de las reglas

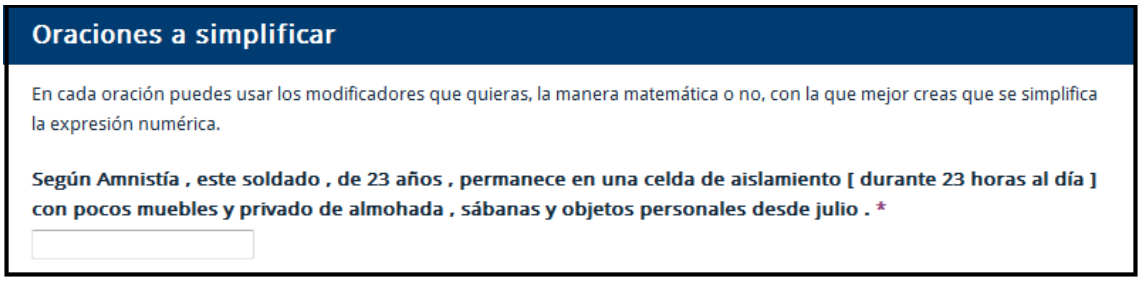

Figura 3: Ejemplo de un parte de la encuesta

En las versiones simplificadas de los textos, un número significativo de estas expresiones numéricas son eliminadas: haciendo el cálculo, menos de la mitad de estas expresiones en los textos originales se han conservado en sus versiones simplificadas. De las expresiones que no se eliminan, la mayoría contienen algún tipo de modificación y en el texto simplificado se presentan de forma diferente a la que aparece en el texto original. También hemos observado un uso variado de modificadores, entre ellos, *más de*, *cerca de*, *casi*, etc.

Ha habido casos en que las expresiones numéricas son eliminadas, en otros casos el número original se redondea cuando una expresión es sustituida por otra, o casos en que el número fue redondeado usando además un modificador añadido que no estaba presente en el texto original. En los trabajos previos (Bautista et al., 2011), (Power y Williams, 2012) ya se sugiere que los modificadores pueden ser una herramienta útil para simplificar una variedad de diferentes expresiones numéricas.

Lo que sigue es un resumen de las operaciones de simplificación más comunes aplicadas a expresiones numéricas en el corpus:

1. Los números en parentésis se eliminan (esta operación ha sido aplicada en un 100 % de los casos en la simplificación manual):

   un millón de francos suizos (unos 770.000 euros) ⇒ *un millón de francos suizos*

2. Los números en letras se sustituyen por números expresados con dígitos:

   nueve millones ⇒ *9 millones*

3. Las grandes cantidades se expresan por medio de una palabra en lugar de dígitos:

   unos 370.000 niños ⇒ *más de 300 mil niños*

4. Grandes números se redondean:

   casi 7.400 millones de euros ⇒ *más de 7000 millones de euros*

5. Se aplica redondeo eliminando puntos decimales:

   1,9 millones de hogares ⇒ *2 millones de casas*[13]

Tras el análisis del corpus, teniendo en mente una futura implementación computacional de las reglas identificadas, se lleva a cabo una encuesta dirigida exclusivamente a la simplificación de expresiones numéricas para observar el uso de modificadores y las estrategias de simplificación

[13] Aquí otro cambio léxico es aplicado: hogar ⇒ casa

aplicadas. Recopilando esta información, podemos completar los resultados obtenidos en el estudio de corpus antes mencionado de cara a la implementación.

4.2 Resultados de la encuesta

Los datos recogidos a partir de la encuesta realizada han sido analizados para identificar las operaciones de simplificación que los participantes han usado para simplificar las expresiones numéricas.

Para cada expresión numérica en una oración dada identificamos todas las operaciones usadas por todos los participantes. Se han identificado un total de 26 operaciones diferentes aplicadas para simplificar las expresiones dadas en la encuesta. Algunos ejemplos son añadir una explicación, calcular el tanto por ciento dado, cambiar de porcentaje a fracción, etc. No todas las operaciones ocurren con suficiente frecuencia como para tenerlas en cuenta en el análisis, por lo que han sido agrupadas dependiendo del tipo de cambio aplicado (por ejemplo si han usado o no modificador) o si la información ha sido eliminada, la cantidad redondeada o la expresión númerica reescrita. Por eso, nos centramos en las operaciones más comunes aplicadas por los participantes.

Como ilustración, veamos el ejemplo de la expresión original *55* en la frase:

> *Amnistía Internacional ha documentado durante 2010 casos de tortura y otros malos tratos en al menos 111 países, juicios injustos en **55**, restricciones a la libertad de expresión en 96 y presos de conciencia encarcelados en 48.*

Las siguientes simplificaciones fueron sugeridas por los sujetos:

- *más de 50*
- *más de la mitad de ellos*
- *la mitad de ellos*
- *55*
- *50*

La expresión simplificada más comúnmente usada fue *más de 50*, donde un modificador es añadido y el número redondeado, aunque con una pequeña pérdida de precisión.

Las observaciones generales que sacamos del análisis de datos obtenidos del cuestionario son las siguientes:

- El número en sí mismo:
  - se deja sin cambios *(26.3 %)*,
  - se redondea *(26.3 % ⇒ más de un 25 %)*,
  - se cambia su forma matemática *(24 % ⇒ casi un cuarto)*,
  - se reescribe en letras *(3 % ⇒ tres por ciento)*,
  - se reescribe en dígitos *(ocho millones ⇒ 8 millones)*
- En ocasiones se pierde precisión de la expresión numérica cuando se sustituye por una versión simplificada. Por ejemplo, *Alrededor de 390.000 personas ⇒ Casi 400.000 personas*
- Si la expresión original no tenía modificador, en ocasiones un modificador es usado en la opción simplificada para tener en cuenta la pérdida de precisión. Por ejemplo, *78 % ⇒ más del 75 %*

En las oraciones presentadas en la encuesta estudiamos, por un lado, las expresiones originales que ya contienen un modificador y, por otro, las que van sin modificador. De las 27 expresiones numéricas originales presentadas en la encuesta, 15 de ellas no tenían modificador mientras que las restantes 12 sí tenían.

En el caso de las 12 expresiones originales con modificadores, en 7 de ellas la operación de simplificación usada más común fue sustituir el modificador original por otro y redondear el número. Esto ocurre con los siguientes modificadores: *al menos* y *casi* son sustituidos por *más de*, mientras que *unos*, *alrededor de* y *cerca de* son sustituidos por *casi*. En 4 expresiones, el modificador original se mantuvo sin cambios, como es el caso de *más de*, *unos* o *unas*, mientras que el número fué redondeado. Hubo sólo un caso donde la expresión numérica original fue completamente reescrita por la mayoría de los participantes en la encuesta y por lo tanto el modificador original se perdió.

Por otro lado, de las 15 expresiones numéricas originales sin modificador, en 8 casos un modificador fue añadido por la mayoría de los participantes; 5 casos continuaron sin modificador (todos ellos debido al hecho de que la simplificación es igual a la original, es decir, no hubo ningún cambio); y en 2 casos la operación más común fue reescribir la expresión numérica original.

Consideramos como casos de reescritura los casos en los que se eliminó la expresión numérica original y se utilizó información textual en su lugar, tal como en el ejemplo siguiente: *durante 23*

horas al día se reescribió como *casi todo el día*. Además, observamos simplificaciones donde un cambio de estrategia de simplificación fue aplicado, como se pueden ver en estos ejemplos: la expresión *26 %* fue simplificada usando una expresión en forma de fracción *una cuarta parte*, y lo mismo fue aplicado en el caso de *34 %*, el cuál fue reescrito como *un tercio*. Los resultados de la encuesta nos hacen ver que el uso de modificadores juega un papel fundamental cuando se simplifican expresiones numéricas.

Nuestros datos muestran que las operaciones más comúnmente aplicadas son añadir un modificador cuando la expresión original no lo tiene ya, y redondear la expresión numérica original, explicado en profundidad en la Sección 4.3.

4.3 Análisis comparativo

Para llevar a cabo un análisis comparativo de los resultados obtenidos en el estudio realizado sobre el corpus y sobre la encuesta, nos centramos en el subconjunto de expresiones númericas usadas en la encuesta y en sus equivalentes en el corpus. Posteriormente, hemos extraído todas las operaciones aplicadas en el proceso de simplificación de las expresiones seleccionadas y comparamos las frecuencias relativas de estas operaciones en el corpus y en la encuesta. Los Cuadros 3 y 4 presentan los resultados. Las filas marcadas corresponden a las operaciones que coinciden en ambos casos.

Operaciones de simplificación	**Número de ExpNum**	**% Uso**
Eliminar ExpNum	12	44.4 %
Eliminar Oración	7	25.9 %
Misma ExpNum	2	7.4 %
Cambiar Modificador + Redondeo	2	7.4 %
Eliminar Modificador + Redondeo	2	7.4 %
Reescribir ExpNum	1	3.7 %
Eliminar Modificador + Mismo número	1	3.7 %
Total	**27**	100 %

Cuadro 3: Operaciones de simplificación obtenidas del análisis del corpus

En los resultados obtenidos del análisis del corpus, más del 50 % de las expresiones numéricas fueron eliminadas, mientras que los resultados de la encuesta sugieren una preferencia por mantener la información a costa de una ligera pérdida de precisión a través de redondeos y compensada por el uso de modificadores. En comparación con la simplificación del corpus, se opta más a menudo por reescribir la información o dejar las expresiones sin modificar, principalmente en los casos de los números grandes como *2.000 millones de dólares*, *más de 20 millones* o *65 millones*.

Operación de simplificación	**Número de ExpNum**	**% Uso**
Añadir Modificador + Redondeo	9	33.3 %
Cambiar Modificador + Redondeo	6	22.2 %
Misma ExpNum	5	18.5 %
Reescribir ExpNum	5	18.5 %
Mantener Modificador + Redondeo	2	7.4 %
Total	**27**	100 %

Cuadro 4: Operaciones de simplificación obtenidas del análisis de la encuesta

En cuanto al uso de los modificadores, los datos recogidos de la encuesta muestran que los modificadores preferidos cuando una expresión numérica se simplifica son: *más de* y *casi*. Estos dos modificadores han sido los más utilizados tanto cuando el modificador de la expresión original se cambia por otro, como cuando el modificador se añade a la expresión ya que inicialmente ésta no contenía ningún tipo de modificador.

Observando las operaciones de simplificación aplicadas por los participantes tanto en la simplificación del corpus como en la encuesta, se puede ver que hay tres operaciones comunes en ambos casos: *Cambiar Modificador + Redondeo, Misma ExpNum* y *Reescribir ExpNum*. La primera y la segunda tienen un uso similar. Obviando los casos de eliminación del corpus, son las dos operaciones más usadas por los expertos en la simplificación de las oraciones con contexto. Y en el caso de la encuesta, sin contar el caso más usado (*Añadir Modificador + Redondeo*), estas operaciones son también muy usadas por los participantes para simplificar las oraciones sin contexto. De ahí que, dependiendo del tipo de la expresión numérica original, una u otra sean usadas para proporcionar una expresión simplificada. En el caso de la última operación, *Reescribir ExpNum*, es mucho más frecuente en el caso de la simplificación de oraciones sin contexto en comparación con el caso de los textos del corpus.

Además, es significativo destacar que de las operaciones no comunes en los dos análisis, en el caso del corpus todas ellas están relacionadas con la eliminación de información (oraciones, expresiones numéricas, modificadores) y en cambio,

en el caso de la encuesta se añade información o se lleva a cabo una transformación de la expresión, manteniendo el modificador pero aplicando un redondeo a la cantidad. Uno de los factores que influye a la hora de detectar tantos casos de eliminación en el caso de la simplificación del corpus, es que cuando se pide simplificar un texto en seguida se asocia con la idea de eliminar información superflua para que así sea más fácil de leer y comprender. Pero esto no siempre es así, ya que la pérdida de información no garantiza un texto más simple. A veces hay que añadir información para ayudar a la lectura y comprensión del texto y entran en juego otros factores, como la frecuencia de uso de las palabras, la ambigüedad y el uso en el contexto de las mismas.

Durante el análisis de las simplificaciones sugeridas por los participantes de la encuesta, detectamos que para algunas de las opciones simplificadas que propusieron el contexto de la expresión numérica dentro de la oración había sido considerado. Veamos por ejemplo en la oración: *Amnistía Internacional ha documentado durante 2010 casos de tortura y otros malos tratos en al menos 111 países, juicios injustos en **55**, restricciones a la libertad de expresión en 96 y presos de conciencia encarcelados en 48.* Para la expresión original **55**, de los casos mostrados en la sección 4.2, podemos observar que dos de las simplificaciones (*más de la mitad de ellos, la mitad de ellos*) han sido propuestas simplificando la expresión original considerando el contexto a nivel de oración y haciendo referencia a los "111 países" nombrados anteriormente. Esto es significativo, porque a pesar de que las oraciones fueron presentadas sin contexto respecto al texto completo, algunas simplificaciones de expresiones numéricas propuestas por los participantes sí que consideraron el contexto a nivel de oración para generar una versión simplificada.

5 Discusión y Conclusiones

Los casos de eliminación, de la oración entera o justo de la expresión numérica en concreto, sólo aparecen en el análisis del corpus. Esto se debe al hecho de que los ejemplos dados en la encuesta eran oraciones individuales sin información añadida, mientras que los ejemplos en el corpus siempre van acompañados por contexto. Por lo tanto, en las oraciones de la encuesta no se producen casos de eliminación de la expresión numérica, y menos de la oración completa, ya que no se daba información añadida de donde aparecía la oración en el texto original.

Además hay que señalar que no se dió como posibilidad a los participantes la opción de eliminar información, solo de simplificar las expresiones que aparecían en cada oración. Estos casos ponen de relieve el papel importante que juega el contexto a la hora de decidir si eliminar o modificar una expresión numérica en una oración.

La simplificación manual del corpus se hizo sabiendo que el lector final sería una persona con discapacidad cognitiva mientras que en la encuesta no se especificó ningún usuario final a quien iban dirigidas las simplificaciones de las oraciones que se presentaban. Por lo tanto, lo que se tiene que decidir es si se debe dar preferencia a la preservación de la información a coste de la precisión, o eliminar la información superflua por completo de un texto que contiene expresiones numéricas.

El corpus que hemos utilizado en este trabajo, ha sido simplificado teniendo en cuenta el contexto y con conocimiento del usuario final a quien iba dirigida la simplificación. Estos dos factores permiten una eliminación selectiva con pérdida muy controlada de información (porque al usuario no le va a servir o porque ya se extrae del contexto).

Dentro del conjunto de operaciones de simplificación identificadas, observamos que hay operaciones comunes a la hora de simplificar las expresiones numéricas teniendo en cuenta el contexto (corpus) y sin tener en cuenta el contexto del texto (encuesta). Lo que demuestra que hay operaciones que, a priori, son más independientes del contexto, y que se aplican en ambos casos, obteniendo una versión simplificada de la expresión numérica que se quiere adaptar.

Es significativo que usando el analizador Freeling seamos capaces de identificar y anotar diferentes tipos y muchos casos distintos de expresiones numéricas, ya que en comparación con otros analizadores basados en aprendizaje automático como, OpenNLP[14], Maltparser[15], Mate-tools[16], que basan su análisis en el corpus que se utiliza para su entrenamiento, y usan la anotación del Penn Treebank POS, en la que sólo se dispone de una única etiqueta para categorías gramaticales (POS) para la información numérica que es *CD*, no pueden dar mayor detalle de qué tipo de expresión numérica ha sido identificada.

Este estudio realizado corrobora las conclusiones previas de los trabajos de Bautista et al.

[14]http://opennlp.apache.org/documentation.html [Último acceso: 20/11/2012]

[15]http://www.maltparser.org/ [Último acceso: 20/11/2012]

[16]http://code.google.com/p/mate-tools/ [Último acceso: 20/11/2012]

(2011) y Power y Williams (2012), sobre el uso de modificadores y el uso de distintas estrategias de simplificación, en este caso para la adaptación de textos en español.

6 Trabajo Futuro

Como parte de nuestro trabajo futuro tenemos la intención de reunir un corpus más rico en expresiones numéricas variadas y repetir el estudio con los editores humanos con el fin de extraer más posibles operaciones de simplificación para otros tipos de expresiones aquí no tratadas, como son por ejemplo el tratamiento de los porcentajes.

Además de esto, tenemos planeado incluir información sobre el usuario final para el que se está simplificando como un factor más a tener en cuenta, ya que las simplificaciones pueden variar dependiendo de para quién se simplifique el texto original. Si se opta por perder precisión, preservarla o eliminar la información que no sea necesaria, tomar estas decisiones en gran medida depende del tipo de lector para el que vaya destinado el texto simplificado.

Desde el punto de vista de eliminación de información, un posible enfoque es utilizar técnicas de resumen automático para desarrollar un clasificador que se pueda emplear como herramienta para la simplificación de textos, y ayude a decidir qué contenido guardar y qué elementos borrar, donde el número de expresiones numéricas se utiliza como un rasgo para crear el clasificador (Drndarević y Saggion, 2012).

El último objetivo de nuestro trabajo es llevar a cabo la implementación de las operaciones detectadas para la simplificación de expresiones numéricas en español, como una categoría específica de expresiones léxicas. Los resultados de los dos análisis realizados se usarán para esta implementación, considerando que algunas expresiones numéricas podrían ser eliminadas dependiendo del contexto y otras sustituidas para hacerlas más accesibles. Para ello tenemos la intención de llevar a cabo un análisis de los datos más profundo y detallado sobre un corpus extenso y obtener así un conjunto de reglas de transformación considerando además las necesidades del usuario final.

Agradecimientos

Queremos agradecer al Dr. Stefan Bott por su ayuda ofrecida con el manejo del analizador Freeling para realizar este trabajo.

Este trabajo ha sido parcialmente financiado por el Gobierno Español a través del Ministerio de Educación y Ciencia (TIN2009-14659-C03-01 Proyecto), Universidad Complutense de Madrid y Banco Santander Central Hispano (GR58/08 Beca de grupo de investigación) y el programa de becas de Formación de Personal de Investigación (FPI).

Este trabajo, en parte, ha sido realizado bajo el proyecto titulado Simplext: un sistema automático para simplificación de textos (Simplext: An automatic system for text simplification), con el número TSI-020302-2010-84 [17]. También queremos agradecer a la financiación del Programa Ramón y Cajal 2009 (RYC-2009-04291), Ministerio de Economía y Competitividad, Secretaría de Estado de Investigación, Desarrollo e Innovación de España.

Bibliografía

Anula, A. 2007. Tipos de textos, complejidad lingüística y facilicitación lectora. En *Actas del Sexto Congreso de Hispanistas de Asia*, páginas 45–61.

Anula, A. 2008. Lecturas adaptadas a la enseñanza del español como L2: variables lingüísticas para la determinación del nivel de legibilidad. En *La evaluación en el aprendizaje y la enseñanza del español como LE/L2, Pastor y Roca (eds.),*, páginas 162–170, Alicante.

Aswani, N., V. Tablan, K. Bontcheva, y H. Cunningham. 2005. Indexing and Querying Linguistic Metadata and Document Content. En *Proceedings of Fifth International Conference on Recent Advances in Natural Language Processing*, Borovets, Bulgaria.

Bautista, S., P. Gervás, y R.I. Madrid. 2009. Feasibility analysis for semiautomatic conversion of text to improve readability. En *The Second International Conference on Information and Communication Technologies and Accessibility*, May 2009.

Bautista, S., R. Hervás, P. Gervás, R. Power, y S. Williams. 2011. How to Make Numerical Information Accessible: Experimental Identification of Simplification Strategies. En *Conference on Human-Computer Interaction*, Lisbon, Portugal.

Biran, O., S. Brody, y N. Elhadad. 2011. Putting it Simply: a Context-Aware Approach to Lexical Simplificaion. En *Proceedings of the ACL.*

[17] http://www.simplext.es [Último acceso: 20/11/2012]

Bott, S. y H. Saggion. 2011. An Unsupervised Alignment Algorithm for Text Simplification Corpus Construction. En *Workshop on Monolingual Text-to-Text Generation*, Portland, USA, June. ACL.

Burstein, J., J. Shore, J. Sabatini, Yong-Won Lee, y M. Ventura. 2007. The automated text adaptation tool. En Candace L. Sidner Tanja Schultz Matthew Stone, y ChengXiang Zhai, editores, *HLT-NAACL (Demonstrations)*, páginas 3–4. The Association for Computational Linguistics.

Candido, Jr., A., E. Maziero, C. Gasperin, Thiago. A. S. Pardo, L. Specia, y Sandra M. Aluisio. 2009. Supporting the adaptation of texts for poor literacy readers: a text simplification editor for brazilian portuguese. En *Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications*, páginas 34–42, Stroudsburg, PA, USA. Association for Computational Linguistics.

Carroll, J., G. Minnen, Y. Canning, S. Devlin, y J. Tait. 1998. Practical Simplification of English Newspaper Text to Assist Aphasic Readers. En *Proceedings of AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, páginas 7–10, Madison, Wisconsin.

Chandrasekar, Raman, Christine Doran, y Bangalore Srinivas. 1996. Motivations and Methods for Text Simplification. En *COLING*, páginas 1041–1044.

De Belder, J., K. Deschacht, y Marie-Francine Moens. 2010. Lexical simpli?cation. En *Proceedings of Itec2010 : 1st International Conference on Interdisciplinary Research on Technology, Education and Communication.*

Deschacht, Koen y Marie-Francine Moens. 2009. Semi-supervised semantic role labeling using the latent words language model. En *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, páginas 21–29, Stroudsburg, PA, USA. Association for Computational Linguistics.

Devlin, S. y G. Unthank. 2006. Helping aphasic people process online information. En *Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility*, Assets '06, páginas 225–226, New York, NY, USA.

Dieckmann, Nathan F., Paul Slovic, y Ellen M. Peters. 2009. The use of narrative evidence and explicit likelihood by decision-makers varying in numeracy. *Risk Analysis*, 29(10).

Drndarević, Biljana y Horacio Saggion. 2012. Reducing text complexity through automatic lexical simplificacion: an empirical study for spanish. *Procesamiento del Lenguaje Natural.*

Inui, K., A. Fujita, T. Takahashi, R. Iida, y T. Iwakura. 2003. Text simplification for reading assistance: A project note. En *In Proceedings of the 2nd International Workshop on Paraphrasing: Paraphrase Acquisition and Applications*, páginas 9–16.

Klebanov, B. B., K. Knight, y D. Marcu. 2004. Text simplification for information-seeking applications. En *On the Move to Meaningful Internet Systems, Lecture Notes in Computer Science*, páginas 735–747.

Lal, P. y S. Ruger. 2002. Extract-based summarization with simplification. En *Proceedings of the ACL 2002 Automatic Summarization.*

Maynard, D., V. Tablan, H. Cunningham, C. Ursu, H. Saggion, K. Bontcheva, y Y. Wilks. 2002. Architectural Elements of Language Engineering Robustness. *Journal of Natural Language Engineering – Special Issue on Robust Methods in Analysis of Natural Language Data*, 8(2/3):257–274.

Medero, J. y M. Ostendorf. 2011. Identifying targets for syntactic simplification. En *In Proceedings of the Workshop on Speech and Language Technology in Education.*

Mishra H, Mishra A, Shiv B. 2011. In praise of vagueness: malleability of vague information as a performance booster. *Psychological Science*, 22(6):733–8, April.

Padró, Ll., M. Collado, S. Reese, M. Lloberes, y I. Castelln. 2010. Freeling 2.1: Five years of open-source language processing tools. En *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, Valletta, Malta.

Peters, Ellen, Judith Hibbard, Paul Slovic, y Nathan Dieckmann. 2007. Numeracy skill and the communication, comprehension, and use of risk-benefit information. *Health Affairs*, 26(3):741–748.

Power, Richard y Sandra Williams. 2012. Generating numerical approximations. *Computational Linguistics*, 38(1).

Qualifications y Curriculum Authority. 2010. Annual report and accounts. Informe técnico, Financial statements.

Quinlan, P. 1992. *The Oxford Psycholinguistic Database*. Oxford University Press.

Saggion, Horacio, Elena Gómez-Martínez, Alberto Anula, Lorena Bourg, y Estaban Etayo. 2011. Text simplification in simplext: Making texts more accessible. En *Proceedings of the Sociedad Española del Procesamiento del Lenguaje Natural*.

Siddharthan, Advaith. 2002. Resolving Attachment and Clause Boundary Amgiguities for Simplifying Relative Clause Constructs. En *Proceedings of the Student Research Workshop, 40th Meeting of the Association for Computacional Linguistics*.

Specia, L. 2010. Translating from Complex to Simplified Sentences. En *9th International Conference on Computational Processing of the Portuguese Language*, páginas 30–39.

Yatskar, M., Pang B., C. Danescu-Niculescu-Mizil, y L. Lee. 2010. For the sake of simplicity: Unsupervised extraction of lexical simplifications from wikipedia. *CoRR*, abs/1008.1986.