UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE INFORMÁTICA

DEPARTAMENTO DE INGENIERÍA DEL SOFTWARE E INTELIGENCIA ARTIFICIAL

TESIS DOCTORAL

Un modelo computacional para la simplificación automática de expresiones numéricas

MEMORIA PARA OPTAR AL GRADO DE DOCTORA
PRESENTADA POR

Susana Bautista Blasco

Directores

Pablo Gervás Gómez-Navarro
Raquel Hervás Ballesteros

Madrid, 2015

Un modelo computacional para la simplificación automática de expresiones numéricas

TESIS DOCTORAL

Susana Bautista Blasco

Departamento de Ingeniería del Software e Inteligencia Artificial
Facultad de Informática
Universidad Complutense de Madrid

Madrid 2015

Un modelo computacional para la simplificación automática de expresiones numéricas

Memoria que presenta para optar al título de Doctora en Informática

Susana Bautista Blasco

Dirigida por

Prof. Dr. D. Pablo Gervás Gómez-Navarro

Prof. Dra. Dña. Raquel Hervás Ballesteros

Departamento de Ingeniería del Software e Inteligencia Artificial

Facultad de Informática

Universidad Complutense de Madrid

Madrid 2015

A mis padres,
a mis hermanos
y a mi sobrino David.

¿Para qué vivimos,
si no es para hacernos la vida
más llevadera unos a otros?
Mary Ann Evans

Agradecimientos

Es mejor perderse que nunca embarcar,
mejor tentarse a dejar de intentar,
aunque ya ves que
no es tan fácil empezar.

Diego Torres

Un día *Pablo Gervás* me dijo que "estando cerca un padre, un hermano, un novio, un amigo o un jefe, las mujeres teníamos claro que ¡la culpa siempre es de él!" ¡Qué grande eres, Pablo! Esta tesis es culpa de él, mi director de tesis, el *Doctor Pablo Gervás*, que desde el primer momento me ha animado y apoyado para seguir luchando por lo que yo quiero.

Mi otro pilar de apoyo ha sido mi otra directora de tesis, la *Doctora Raquel Hervás* que siempre ha confiado en mí, en mi trabajo, ha sabido animarme en los momentos difíciles, y juntas hemos trabajado mucho y muy bien. Ambos han tenido siempre la palabra adecuada para seguir animándome en el duro camino que es una tesis doctoral. En ellos siempre he tenido un apoyo incondicional y no tengo suficientes palabras de agradecimiento, pero ambos saben lo que significan para mí.

Quiero agradecer a la *Doctora Sandra Williams* con la que realicé mi primera estancia de investigación en la Open University, en Milton Keynes, Reino Unido. Con ella aprendí y crecí mucho a nivel profesional. Esta estancia fue una experiencia muy buena en mi vida.

Además, realicé otra estancia en Barcelona, en la Universitat Pompeu Fabra, con el *Doctor Horacio Saggion*, para el que sólo tengo palabras de agradecimiento. Este tiempo fue muy fructífero a nivel profesional y personal, una etapa muy importante de mi vida y de la tesis.

A los miembros de mi grupo de investigación *Natural Interaction based on Language* (NIL): *Alberto*, *Fede*, *Carlos*, *Virginia* y *Gonzalo*, donde he crecido como persona y como investigadora y donde me han ayudado a ir superando los distintos obstáculos que me iba encontrando. Gracias por vuestra paciencia y comprensión. Una mención especial a mi amigo *Gonzalo*, gracias por tu ayuda incondicional, por los cafés que nos hemos tomado y por las risas que hemos compartido.

Agradecer al Departamento de Ingeniería del Software e Inteligencia Artificial (DISIA), destacando al director del mismo, *Dr. Luis Hernández*, que siempre ha confiado en mí, ayudándome en todo momento y brindándome la oportunidad de encontrar mi hueco en el departamento. Quiero nombrar a otras personas importantes para mí del departamento, que han estado ahí cuando más lo he necesitado: *Guille*, *Eva*, *Belén*, *Antonio*, *Javi*, *Almu* y *Lourdes*. Gracias por vuestras palabras de ánimo, por vuestro apoyo y por confiar siempre en mí. A los miembros del grupo TALN de la Universitat Pompeu Fabra donde realicé una de mis estancias, y en especial a *Roberto*, *Biljana*, *Mireia*, *Alicia*, *Luz* y *Simon*, fuísteis unos grandes compañeros.

Una mención especial a mis compañeros de despacho del 218, *Ana* y *Javi*, sin vosotros no tengo muy claro dónde estaría yo ahora. Ellos vivieron mis primeros pasos en el departamento, mis dudas, mis preocupaciones y mis alegrías del comienzo de esta aventura. Juntos hemos compartido mucho, dentro y fuera de la facultad. Nuestro despacho era, es y será una referencia para el resto de la Facultad. Gracias por quererme tanto.

A mis compañeros de despacho 420Bis: a *Kiko*, porque siempre sabe sacarme una sonrisa aún en el peor de mis días. A *Dani*, del que admiro su temple y paciencia, ¡gracias por estar ahí, genio! A *Iván*, con el que me he reído mucho a pesar de su seriedad en el trabajo. A *Carlos*, con el que me entiendo sin hablarnos. Gracias chicos por cuidarme tanto.

A la Facultad de Informática, en la que he pasado los últimos 14 años de mi vida. He vivido distintas etapas en la facultad de manera muy diferente, cuando no teníamos edificio, cuando se construyó, cuando nos mudamos, de estudiante, de becaria, contratada, y en cada momento me he ido encontrando con gente maravillosa y otra no tanto, pero que han hecho que esto funcione, aunque parezca mentira. Gracias a *Milagros Fernández*, mi profesora de Tecnología de Computadores de 2º de carrera, por su entusiasmo y apoyo. Gracias a *Román Hermida* porque durante su etapa como Decano, me brindó el apoyo que necesité en uno de los momentos más difíciles de mi carrera de ingeniería. A *Daniel Mozos*, a *Narciso Martí* y a *José Antonio Macarrón* por su cercanía y confianza. A la gente de la cafetería, *Andrés*, *Richi*, *Sánchez* y *Manolo*, porque durante mis años en la facultad, han sido los responsables de mi alimentación, hemos compartido risas y muchas fiestas. Gracias por mimarme tanto.

Al *Doctor Carlos León*, con el que he crecido como persona y como investigadora, sin él no estoy segura de si yo estaría aquí. Gracias Doctor por tus palabras sabias en cada momento, por tu confianza y por tu apoyo incondicional, gracias por quererme tanto. Al *Doctor Miguel Ballesteros*, por tu escucha y ayuda, gracias por nuestras conversaciones.

Quiero agradecer a cada uno de los miembros de la *"Cena del Anillo"* con los que he descubierto joyas gastronómicas y siempre me han apoyado en todo momento. Hemos compartido muy buenos momentos juntos. Agradecer

también a los componentes del grupo *"Birras"*, por los buenos reencuentros y celebraciones en la cafetería.

A mis amigas "las kukis", *Laura*,*Vir* y *Mari Cruz*, gracias por quererme, respetarme y apoyarme siempre, es un lujo teneros como amigas. Hemos compartido muchos momentos juntas y todos los que nos quedan cuando sea *Doctora*.

A mi amiga *Ana Mellado*, a la que un día me unió la Ingeniería de Informática. Pese a las dificultades de la vida, siempre encuentro un apoyo en ella, una escucha incondional y sólo puedo dar las gracias por tenerla a mi lado. Junto a Ana, mi paso por la facultad me ha brindado la oportunidad de tener grandes amigos informáticos: *Dani*, *Manu*, *Murgui*, *Carlos*, *Patxi*, *Germán*, *María*, *Álvaro*, *Rebe*, *el Abu*, *Rubén*, *Pedro*, gracias por confiar siempre en mí y quererme como soy.

A "La Asunción" a través de la cual he crecido como persona, en un camino de vida acompañada por muchas grandes y bellas personas, y que me ha brindado la oportunidad de tener amigos repartidos por toda España: Gijón, Málaga, Ronda, Dalías, Tenerife, Cartagena, Algete, Alcobendas... *Paula*, *María José*, *Marta*, *Gloria*, *Rosi*, *Juanjo*, *Rubén*, *Pedro*, *Patri*, *Isa*, *Fali*, *Lorena*, *Javi*, *Yurena*, *Dani*, *Cecilia*, *Miry*, *Pili*, *Richi*, sin vosotros no sería la persona que soy hoy, gracias por formar parte de mi vida.

Quiero agradecer su esfuerzo y dedicación a todos los participantes de los experimentos que he realizado a lo largo de esta tesis. Gracias a vosotros he podido desarrollar y evaluar los sistemas que presento en este trabajo. Gracias en especial a todos los profesores que han participado, y en especial a *Víctor*, por su ayuda prestada en la evaluación del sistema en español. Gracias a *Sandra* por ofrecerse a ayudarme a darle un toque de diseño a la tesis. A *Scott*, a *Sergio*, a *Luis* y a *Leti* por su ayuda en la corrección de este documento. Además, quiero agradecer a *Ricardo García* por sus enseñanzas y su ayuda en los estudios estadísticos realizados a lo largo de esta tesis.

Además, quiero dar las gracias a los tres revisores que han generado los informes internos y europeos, ya que gracias a sus comentarios y revisiones he podido mejorar el trabajo que presento en mi tesis. Gracias al Doctor *Gonzalo Méndez* de la Universidad Complutense de Madrid, al Doctor *Klaus Miesenberger* de la Universidad de Linz, Austria y a la Doctora *Barbara Arfé*, de la Universidad de Padua, Italia.

A mi madre *Pepi*, por su confianza, por sus consejos, por su escucha y por su apoyo incondicional. Sin ti, no sé que hubiera sido de mí. Gracias mamá por estar siempre ahí, superando las barreras de tiempo y distancia. Gracias a mi padre *Jerónimo* y a mis hermanos *Jero*, *Floren* y *Jose*, por apoyar y respetar mis decisiones. Ha sido un camino largo, pero siempre habéis estado ahí. A mi sobrino David, para que vea que el esfuerzo y la constancia, tienen su recompensa.

Y para terminar, a otros muchos que no he nombrado aquí, pero que

saben que han estado en distintas etapas de mi vida, y que de una manera u otra han hecho posible que hoy pueda estar escribiendo estas líneas y sea la persona que soy.

Tras la defensa de esta Tesis Doctoral, cierro una etapa y comienzo a escribir un nuevo capítulo de mi vida. Gracias a cada uno por formar parte de ella.

Susana Bautista Blasco

Resumen

La manera en la que se escribe o se presenta la información escrita puede provocar problemas de acceso a la información a un gran número de personas que tienen dificultades para la comprensión de textos. Estos problemas pueden ser debidos a diversos factores como por ejemplo haber tenido un acceso limitado a la formación, estar en riesgo de exclusión social o tener alguna discapacidad cognitiva. En concreto, existen colectivos específicos como las personas sordas, autistas, personas con trastornos del lenguaje como afasia o dislexia, personas mayores o personas que están aprendiendo otro idioma, que tienen problemas con la lectura. Con el objetivo de hacer la información accesible para todos hay que tener en cuenta la diversidad de las personas que van a acceder a ella.

El trabajo presentado en esta tesis se enmarca dentro de la línea de investigación de la simplificación automática de textos y en concreto en el tratamiento de la información numérica. La simplificación de textos tiene como objetivo transformar un texto en otro similar que sea más fácil de leer. Para ello, hay que identificar qué provoca dificultad en los lectores y definir diferentes transformaciones, principalmente dirigidas a construcciones sintácticas y léxicas, que se aplican al texto original y generar una versión simplificada del mismo.

En primer lugar, se han revisado las distintas aproximaciones de simplificación automática de textos implementadas en el área, haciendo especial hincapié en aquellas que tratan información numérica. Con todo esto, diseñamos e implementamos un modelo para la simplificación de textos centrado en expresiones numéricas. Para ello, presentamos las bases teóricas para la simplificación de textos junto con el modelo, y mostramos la identificación experimental de las estrategias de simplificación de expresiones numéricas realizada para decidir qué tipo de transformaciones hay que implementar para nuestras aproximaciones automáticas. Finalmente, presentamos el desarrollo e implementación de dos sistemas de simplificación de expresiones numéricas en inglés y en español que siguen el modelo presentado y utilizan las pautas descubiertas en los casos de estudio experimentales llevados a cabo. Para ambos sistemas se realiza una evaluación con expertos que nos permite evaluar la salida de nuestros sistemas.

Resumen en Lectura Fácil

Accedemos a información que a veces es difícil de leer y entender. Hay personas que tienen problemas por diversas razones: están aprendiendo otro idioma, personas mayores o personas con discapacidad. Queremos hacer que la información sea accesible para todos, que cualquier persona pueda leer.

La simplificación de textos sirve para hacer textos más fáciles de leer. El ordenador hace de forma automática transformaciones en el texto más rápido que una persona.

El trabajo que presentamos se centra en un tipo de transformaciones concretas. Queremos simplificar las expresiones numéricas que estén en el texto. Porque a veces la información numérica es difícil de entender y queremos simplificarla.

Hemos definido un modelo para realizar la simplificación de textos en el ordenador. Para ello, les preguntamos a las personas expertas para que nos dijeran qué tipo de transformaciones aplican ellos para simplificar un texto y así enseñar al ordenador.

Hemos implementado dos sistemas de simplificación automática de expresiones numéricas para textos en inglés y en español. Las personas expertas han evaluado nuestros sistemas para comprobar cómo realizan las simplificaciones de manera automática.

[Este resumen en lectura fácil ha sido generado manualmente por la autora de esta tesis.]

Abstract

The way of writing or presenting information can exclude many people, especially those who have problems to read and write or to understand. There are different factors as for example limited cultural education, people have cognitive problems or another disability, people with social problems or people whose mother tongue is not the official language of their adoption country which can cause problems. In particular, there are specific groups like deaf people, autistic people, elderly or people with language disorders such as aphasia or dyslexia, who have problems when they access information. In order to make information accessible to all people, we must keep in mind the diversity of the people who will access it.

The work presented in this thesis is included within the research of automatic text simplification and particularly in the treatment of numerical information. Text simplification aims to transform a text into a similar text that is easier to read. To do this, one has to identify what causes difficulties to readers and define different transformations, mainly aimed at syntactic and lexical constructions that can be applied to the original text to generate a simplified version.

First, we reviewed related approaches to automatic text simplification implemented in the area, with particular emphasis on those dealing with numerical information. Our proposed work on automatic simplification of numerical expressions, is a computational implementation based on a generic model of the process. To this end, we present the theoretical bases for text simplification along with a generic model, and show the experimental identification of simplification strategies on numerical expressions to decide what kind of changes need to be implemented for our automatic approaches. Finally, we present the development and implementation of two systems to simplify numerical expressions in English and Spanish that follow the generic model and use the simplification strategies identified in the experimental studies. For both systems an evaluation with experts has been carried out.

Abstract easy-to-read

We access information that is sometimes difficult to read and understand. Some people have problems for several reasons: they are learning another language, they are elderly or they have special needs. We want to make information accessible to all, so that anyone can read.

Text simplification is used for making text easy-to-read. The computer automatically makes changes to the text faster than a person, making the final version easier to understand.

The present thesis focuses on a specific type of transformations. We want to simplify numerical expressions that are in the text. Because sometimes the numerical information is hard to understand.

We defined a model for text simplification on the computer. To do this, we asked the experts in order to know what kind of transformations they apply to simplify the text and thus teach the computer.

We have implemented two systems to simplify numerical expressions in English and Spanish. The experts have evaluated our systems to analyze the output generated.

[This abstract has been generated manually by the author]

Índice

Índice de figuras

Índice de Tablas

Parte I

Un modelo computacional para la simplificación automática de expresiones numéricas

Capítulo 1

Introducción

1.1. Introducción

Vivimos en la Sociedad de las TIC (Tecnologías de la Información y la Comunicación), expresión que va siendo cada vez más habitual, y que se entiende como el conjunto de tecnologías, recursos, procedimientos y técnicas usadas en el procesamiento, acceso, almacenamiento y transmisión de información presentada en diferentes formatos. Como resultado de esta sociedad hay una tendencia a digitalizar todo tipo de información, noticias, recetas, informes, documentos oficiales, normativas o guías, con el objetivo de hacer la información más accesible a los usuarios. Sin embargo, los estudios realizados muestran que todavía estamos lejos del ideal de una sociedad uniformemente digitalizada donde la información sea accesible a todo el mundo.

El modo en que se escribe o se presenta la información escrita puede excluir a un gran número de personas cuyo nivel de habilidades lectoras les hace tener problemas en la comprensión de textos. Existen diversos factores por los que estas habilidades se pueden ver afectadas como, por ejemplo, haber tenido un acceso limitado a la formación, tener problemas sociales o tener alguna discapacidad cognitiva. Además, existen colectivos concretos como las personas sordas, autistas, personas con trastornos del lenguaje como afasia o dislexia, personas que están aprendiendo otro idioma o personas mayores, que tienen problemas específicos con la lectura. A la hora de presentar la información escrita hay que tener en cuenta la diversidad de las personas que van a acceder a ella y hacer que sea lo más fácilmente accesible para todos.

Las Normas Uniformes sobre la Igualdad de Oportunidades para Personas con Discapacidad de Naciones Unidas (UN, 1994) requieren a los gobiernos que hagan accesibles todos sus servicios públicos de información y documentación a los diferentes grupos de personas con discapacidad, promoviendo a su vez entre los medios de comunicación - televisión, radio y prensa - que sus servicios sean accesibles a todas las personas. Ya que el acceso a la informa-

ción para el desarrollo social y cultural es un derecho fundamental para la igualdad entre las personas. La problemática social ante la que nos encontramos es que existe dificultad a la hora de acceder a la información, ya que esta información se presenta de un modo que dificulta la lectura y comprensión del contenido de la información para distintos colectivos de la sociedad.

La primera solución para esta problemática es la simplificación de la información de manera manual para adaptarla según las dificultades de los usuarios finales a los que va dirigida. Sin embargo, la simplificación manual es demasiado lenta y tediosa para ser eficiente a la hora de producir el material deseado. Al ritmo que avanza la información en la era tecnológica en la que las noticias vuelan por la red, y en tiempo real se distribuyen por diversos medios, no es factible realizar una simplificación manual de la información. Por eso, diversos intentos por automatizar parte de este proceso de simplificación han sido puestos en marcha centrándose en las diferentes transformaciones que se pueden aplicar en el proceso de simplificación de un texto.

La simplificación automática de textos es una tarea relativamente nueva dentro del Procesamiento del Lenguaje Natural. El objetivo de la simplificación de textos es transformar un texto en otro equivalente que sea más fácil de entender para unos usuarios finales determinados. Para llevar a cabo esta tarea, hay que identificar lo que provoca esa dificultad en los lectores y definir distintas transformaciones, principalmente dirigidas a construcciones sintácticas y léxicas, que puedan ser aplicadas al texto original para generar una versión simplificada del mismo.

Los usuarios a los que van dirigidos los textos que se obtienen tras el proceso de simplificación poseen características muy distintas que divergen a la hora de realizar una adaptación de los textos originales. Cuando hablamos de adaptación de contenidos, nos referimos a la transformación de diferentes contenidos que presentan dificultades para el usuario final. Las dos cuestiones principales son qué adaptar y cómo adaptarlo. La primera cuestión busca los elementos a adaptar con el fin de utilizar correctamente el contenido dado. En cuanto a la cuestión de cómo llevar a cabo la adaptación, depende de las características de los usuarios considerados para realizar la adaptación. La adaptación de contenidos se realiza de una manera u otra dependiendo del usuario final. Las habilidades lectoras y el nivel de comprensión de un texto se ven afectados por muchos factores externos que influyen en la persona como barreras sociales como la pobreza o la falta de formación cultural o de acceso a tecnologías avanzadas. A las personas con dificultades va dirigida esta propuesta de resolver una problemática social que con el crecimiento de la información digital, cada vez va aumentando más y necesita soluciones en tiempo real.

Un caso concreto de información que crea dificultades a los lectores es la información numérica. Muchas veces, la información a la que accedemos

viene representada en forma de expresiones numéricas como por ejemplo datos económicos, estadísticos, demográficos, información numérica de una receta, de una noticia o de un informe. Estas expresiones numéricas pueden causar problemas de comprensión para muchas personas por diversas razones, bien porque tengan discapacidad o baja formación académica.

Un estudio realizado por el Gobierno de Reino Unido en 2011 como parte del *Programme for the International Assessment of Adult Competencies* (PIAAC) estimó que 7,5 millones de adultos (el 22 % de la población) estaban en el Nivel 2 o por debajo en matemáticas y no serían capaces de conseguir el grado C (equivalente a un Aprobado en España) en el examen de matemáticas correspondiente a los estudiantes de 16 años (Williams et al., 2003), (Williams et al., 2012), (Miller y Lewis, 2012) . Aproximadamente dos de cada cinco personas (un 36 %) dijo que sus habilidades matemáticas eran muy débiles y que a veces les causaban problemas en su vida diaria como al pagar las facturas de la casa y entender sus nóminas. Otras áreas más comunes donde la gente se sentía perdida eran en la medición y en el peso (en administración de dosis de medicamentos, medidas de ingredientes en las recetas de cocina, etc.) y por supuesto, en la comprensión de los datos estadísticos que se presentan en los medios de información.

En España, en el último informe del Programa Internacional para la Evaluación de la Competencia de los Adultos (PIAAC, por sus siglas en inglés), más conocido como el informe PISA[1] para adultos, se evaluó el rendimiento en comprensión lectora y en comprensión matemática entre la población de 16 a 65 años. Sólo uno de cada tres españoles sabe leer un texto largo o comparar ofertas. En España, el 71,7 % de los adultos pueden realizar con soltura tareas lectoras y de comprensión de textos sencillos. En comprensión matemática, sólo el 68,6 % de los adultos son capaces de realizar cálculos matemáticos sencillos y tan sólo el 24,5 % es capaz de interpretar estadísticas, gráficas o resolver problemas en pasos. Según el estudio, la gran mayoría de los españoles tienen dificultades para extraer información matemática de situaciones reales, como comparar paquetes de ofertas turísticas, calcular el precio final de una compra con descuentos e interpretar gráficos y estadísticas como los que aparecen, por ejemplo, en los recibos de la luz.

Frente a esta realidad, el objetivo de este trabajo es realizar la simplificación automática de expresiones numéricas presentes en los textos. Ya hemos comentado que la simplificación manual no es efectiva debido al ritmo que cambia la información en la sociedad en la que vivimos. La forma en la que se presenta la información numérica puede causar problemas a la hora de leer y comprender un texto. La simplificación automática llevada a cabo en el trabajo que presentamos en esta tesis está basada en las conclusiones extraídas de un estudio empírico desarrollado con expertos. La adaptación de la información no es un proceso sencillo pero sí claramente necesario.

[1] http://www.mecd.gob.es/inee/estudios/piaac.html

1.2. Motivación

Dentro de la simplificación automática de textos nos centramos en un tipo concreto de información para adaptarla y facilitar así su lectura y comprensión. En nuestro trabajo, la información elegida es la información numérica, que tal y como recogen diversos estudios e informes, es un tipo de información que causa dificultades en diversos colectivos de la sociedad.

Como ejemplo de este tipo de problemática tenemos las noticias de prensa que diariamente se publican, a través de las cuales se presenta todo tipo de información. Podemos observar que muchas de ellas contienen información numérica y la forma en la que ésta es presentada afecta a la lectura y comprensión del texto.

En nuestro trabajo consideramos *expresiones numéricas* a las expresiones que denotan cantidades, opcionalmente acompañadas por un modificador numérico, como es el caso de *más de un cuarto* o *casi el 97 %*, donde *más de* y *casi* juegan el papel de modificador. Este tipo de expresiones son muy frecuentes en los textos informativos cargados de información numérica.

A continuación tenemos un ejemplo de texto de una noticia de prensa, tomada de la Agencia de Noticias Servimedia[2], y pongamos atención en el número y variedad de expresiones numéricas usadas (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 Humanitarios, 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 país**. *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, ya que lo forman seis frases, contando con el título, nos encontramos con 12 expresiones numéricas diferentes, lo

[2] http://www.servimedia.es

que supone una media de 2 expresiones numéricas por oración. Eso incluye expresiones con cantidades representadas en distintos formatos como son fracciones, porcentajes, valores exactos, con decimales o ratios. Además se usan diferentes unidades de medida y modificadores numéricos como son *más de*, *alrededor* o *casi*. Tal carga de información, así como la variedad de expresiones numéricas, pueden afectar a la comprensión del lector del texto y evitar el descubrimiento de las relaciones causa-efecto de los acontecimientos más importantes presentados en el artículo de prensa.

1.3. Objetivos

El acceso a una información accesible para todos es nuestro interés fundamental en este trabajo, y en particular, el caso de acceso a la información numérica. Proponemos un modelo genérico para llevar a cabo el proceso automático de simplificación de textos. A partir de este modelo genérico, nos centramos en la simplificación de un tipo especial de información que es la información numérica. El objetivo es hacer la información numérica más accesible reescribiendo las expresiones numéricas difíciles de una manera más simple. Para ello proponemos unas etapas específicas dentro del modelo genérico para llevar a cabo esta tarea. Esto requiere un conjunto de estrategias de reescritura que produzcan expresiones que sean lingüísticamente correctas, más fáciles de entender que las originales y lo más cercanas posible al sentido original.

A continuación, enumeramos más detalladamente los principales objetivos del presente trabajo de tesis:

1. Explorar el área de simplificación de textos, centrando el foco de atención en un tipo especial de información: la información numérica.
2. Presentar un modelo específico para simplificar expresiones numéricas englobado en el modelo genérico de simplificación automática de textos.
3. Realizar un estudio empírico para identificar estrategias de simplificación de información numérica.
4. Desarrollar e implementar distintos sistemas de simplificación de información numérica, siguiendo el modelo anteriormente propuesto para distintos lenguajes.
5. Evaluar la salida de los sistemas implementados.

1.4. Estructura de la tesis

El presente trabajo de investigación está estructurado en siete capítulos, siendo el primero de ellos esta introducción. A continuación se describen el

resto de capítulos.

- **Capítulo 2: Trabajo relacionado.** Este capítulo presenta las principales áreas de investigación relacionadas con este trabajo. Partiendo de la tarea de simplificar textos, nos centramos en la simplificación de expresiones numéricas. Presentamos los procesos de lectura y razonamiento matemático, junto con las tareas principales en la simplificación de textos. Continuamos con las prácticas existentes de simplificación manual de textos y presentamos las distintas aproximaciones de simplificación automática de textos desarrolladas hasta el momento de presentación de este trabajo.

- **Capítulo 3: Herramientas y recursos.** En este capítulo se presenta el corpus como recurso de simplificación, las distintas herramientas de análisis de texto y las herramientas específicas para el tratamiento de expresiones numéricas.

- **Capítulo 4: Bases teóricas para la simplificación de textos centrada en expresiones numéricas.** Este capítulo muestra la descripción y etapas del modelo genérico para la simplificación de textos, así como el modelo específico para la simplificación de expresiones numéricas. Además, presentamos la identificación de estrategias de simplificación de expresiones numéricas para inglés y para español llevadas a cabo siguiendo diferentes metodologías.

- **Capítulo 5: Sistemas de simplificación de expresiones numéricas.** En este capítulo se presentan los sistemas implementados para la simplificación de expresiones numéricas en inglés y en español, junto con su correspondiente evaluación.

- **Capítulo 6: Discusión.** En este capítulo se discuten el planteamiento del trabajo, el modelo genérico como una abstracción de la práctica existente, la identificación experimental realizada, los sistemas implementados y la interpretación de las expresiones numéricas.

- **Capítulo 7: Conclusiones y trabajo futuro.** Este capítulo muestra las conclusiones extraídas de esta tesis y presentamos las distintas líneas de trabajo futuro que han surgido a lo largo de este trabajo de investigación.

Resumen y conclusiones

En el presente capítulo hemos hecho una introducción a la problemática ante la que nos encontramos, la dificultad de leer y comprender la información escrita a la que accedemos y la necesidad de simplificar dichos textos

para que sean accesibles al mayor número posible de personas. Hemos presentado la motivación del trabajo de esta tesis, centrando la investigación en la simplificación de la información numérica presente en los textos. Para concluir, hemos definido nuestros objetivos de trabajo y la estructura de la presente memoria.

En el siguiente capítulo se lleva a cabo una revisión del trabajo relacionado en el área de investigación en el que se enmarca esta tesis. Presentamos la tarea de simplificar un texto, centrándonos en la simplificación de información numérica. Mostramos los procesos de lectura y razonamiento matemático necesarios para acceder a la información escrita. Revisamos las prácticas existentes de simplificación manual de textos y mostramos las distintas aproximaciones de simplificación automática de textos.

Capítulo 2

Trabajo relacionado

El proceso de simplificación de textos nace de la necesidad de adaptar los contenidos textuales para aquellas personas que tienen dificultades a la hora de comprender un texto de manera que se sientan integradas en la sociedad, ya que el acceso a la información es un derecho fundamental. La simplificación de textos consiste en la transformación de un texto en otro equivalente que es más fácil de leer y entender. El objetivo es conseguir textos más accesibles, más atractivos y comunicativos para que sean interesantes y fomenten la lectura en las personas que tienen dificultades. El acceso a la lectura es una necesidad social y un derecho reconocido, y leer es un placer que permite compartir ideas, pensamientos y experiencias.

Un 30 % de la población tiene dificultades lectoras que pueden ser debidas a distintos factores que hacen que necesiten una versión simplificada del texto para acceder a la información. Entre estos factores están las dificultades interculturales, la complejidad de los textos ante los que nos encontramos y los aspectos cognitivos del lector. Entre las personas que necesitan una adaptación del texto original para que puedan llegar a comprender el contenido del mismo están las personas con discapacidad intelectual, las personas mayores, las personas que están aprendiendo otra lengua distinta a la suya y un rango amplio de personas con necesidades educativas especiales (autistas, personas con afasia, personas con dislexia y personas con déficit de atención).

A la hora de comunicarnos a través de textos escritos lo importante es usar expresiones simples, claras y directas que permitan una mejor comprensión de los textos, consiguiendo una comunicación más eficaz y cercana con el público destinatario, dando un paso más a favor de la inclusión social. Realizando distintas operaciones a nivel léxico y a nivel sintáctico se disminuye la complejidad lingüística, consiguiendo un texto simplificado para un usuario específico y en un idioma concreto.

En este capítulo presentamos la tarea de simplificar un texto y las operaciones principales que se destacan en ella. A continuación, se realiza una revisión de las propuestas de simplificación manual desarrolladas hasta hoy

y de las aproximaciones de simplificación automática que se han implementado en este área. Entre ellas, prestamos especial atención al tratamiento de la información numérica debido al ámbito de trabajo de esta tesis.

2.1. La tarea de simplificar un texto

La necesidad de simplificar un texto aparece por los problemas que surgen al acceder a la información que nos llega, principalmente a la información escrita. Entre los problemas ante los que nos encontramos podemos destacar el uso de un lenguaje complicado y no directo, expresando más de una idea por frase, lo que dificulta la lectura y comprensión del texto. Además, el uso de palabras poco frecuentes, de estructuras sintácticas complicadas, de expresiones numéricas difíciles de entender, tecnicismos, abreviaturas y palabras de otras lenguas pueden dificultar la lectura y comprensión del texto.

Desde un punto de vista lingüístico, la complejidad lingüística de un texto se puede medir a partir de tres dimensiones de análisis: la complejidad léxica (frecuencia y densidad léxica), la complejidad de los segmentos discursivos (longitud y naturaleza) y la complejidad estructural de las oraciones (Anula, 2007). Entendemos como *segmento discursivo* "una oración, una frase, un grupo de oraciones que está comprendido entre dos pausas ortográficas que confieren al segmento independencia sintáctica" (Anula, 2008). Controlando estos factores de complejidad es posible diseñar estrategias de simplificación de los textos para garantizar una correcta comprensión. A partir de los estudios realizados en los trabajos de Anula, se sabe que la frecuencia de uso de las palabras condiciona la comprensión lectora de un texto: cuantas más palabras poco usadas contenga el texto, más difícil será su comprensión. Además, los segmentos más largos y múltiples (con más de una cláusula u oración) presentan mayor complejidad y peor será su comprensión, y cuanto mayor presencia de cláusulas subordinadas haya (sustantiva, adjetiva y adverbial), mayor dificultad presenta la comprensión lectora. Con todo esto, Anula plantea unas técnicas de graduación de la complejidad lingüística de los textos: revisar la frecuencia de las palabras, comprobar la densidad léxica y el nivel de redundancia, reescribir los segmentos recursivos mayores de diez o quince palabras y analizar la complejidad de los segmentos discursivos múltiples, separándolos en segmentos unitarios.

Desde un punto de vista computacional, existen diferentes métricas para medir la fácil lectura de un texto, basadas en ecuaciones matemáticas cuyas medidas correlativas son los elementos de escritura, como el número de pronombres personales en el texto, el número de sílabas por palabras o el número de palabras por oración en el texto. Están disponibles para el inglés[1] y para el castellano (Moro et al., 1993). El problema que presentan estas métricas es que no tienen en cuenta el significado de las palabras ni de las

[1] http://w-shadow.com/blog/2009/04/28/calculating-readability-metrics-in-php/

frases, por lo que no sirven para evaluar ni la corrección lingüística ni el contenido informativo del texto.

Cabe destacar que hay diferentes iniciativas que proponen pautas que pueden ayudar cuando se reescribe un texto para hacerlo mas comprensible. Algunas de ellas son *Plain Language*[2], las *Directrices Europeas para generar Información de Fácil Lectura* (Freyhoff et al., 1998) y las últimas pautas para el *Contenido Accesible en la Web* (WCAG 2.0)[3], con un gran número de recomendaciones para hacer el contenido web más accesible.

El proceso de simplificar textos generalmente se realiza a mano y consume mucho tiempo y esfuerzo. Además, al ritmo que cambian los contenidos digitales en nuestra sociedad, es muy difícil mantenerse al día en todos los medios como noticias, redes sociales o foros. Aún así existen diversos recursos donde se han creado contenidos simplificados manualmente a partir de material ya existente, como es el caso de *Simple English Wikipedia*[4], donde se han simplificado 104.861 artículos de la *Wikipedia*, de una manera más simple, o el portal de Noticias Fácil en español de Discapnet[5], donde las noticias de prensa diariamente son adaptadas manualmente. Esta misma idea es llevada a cabo por el portal de noticias sueco *8Sidor*[6]. A nivel de la Unión Europea, existe el portal *e-include*[7] que representa la voz de más de 60 organizaciones para personas con discapacidad intelectual y sus familias a través de toda Europa.

2.1.1. Simplificación de información numérica

El informe de las Naciones Unidas de las Normas Uniformes sobre la Igualdad de Oportunidades para Personas con Discapacidad (UN, 1994) recomienda que la información pública debería ser accesible para la más amplia población posible. La información se presenta muchas veces en forma de expresiones numéricas (por ejemplo, estadísticas económicas, resultados electorales, datos demográficos, datos del paro...) que presentan problemas de comprensión para muchas personas, entre ellas personas no nativas, con baja formación o con problemas de inclusión o enfermedades o transtornos mentales.

Frente a esta problemática social surge la necesidad de adaptar las expresiones numéricas con las que se presenta la información a la población, simplificando expresiones numéricas más complicadas de entender por otras que sean más simples y más fáciles de entender. Para nuestra propuesta necesitamos un conjunto de estrategias de reescritura que produzcan expre-

[2] http://www.plainlanguage.gov
[3] http://www.w3.org/TR/WCAG/
[4] http://simple.wikipedia.org
[5] http://www.noticiasfacil.es/ES/Paginas/index.aspx
[6] http://www.8sidor.se/
[7] http://www.e-include.eu/

siones que sean lingüísticamente correctas, más fáciles de entender que las originales, y tan cerca como sea posible del significado de las expresiones inicialmente usadas. Por ejemplo, *50.9 %* podría ser reescrito por *un poco más de la mitad*. En esta reescritura, los modificadores juegan un importante papel, indicando que se ha llevado a cabo una aproximación a la cantidad original. La simplificación de expresiones numéricas en algunos casos implica una pérdida de precisión, algo que no es necesariamente malo por varias razones:

1. La pérdida de precisión puede ser salvada lingüísticamente, usando modificadores como *casi* o *más de* que simplifican la expresión original sin perder su significado, tan sólo perdiendo precisión que muchas veces no es necesaria.

2. Krifka (2002) argumentó que los escritores y los hablantes con mucha frecuencia aproximan información numérica y los lectores y receptores pueden fácilmente reconocer la información aproximada, incluso cuando no hay modificador que acompañe a la expresión numérica. Esto ocurre especialmente cuando los números son redondeados. Por ejemplo, en *la distancia de Oxford a Cambridge es de 100 millas*, está claro que 100 millas es una aproximación, ya que la distancia real es 100,48 millas.

3. Dubois (1987) verificó en un estudio empírico el uso de modificadores en las expresiones numéricas, observando las presentaciones científicas donde los resultados que se presentaban se aproximaban a la hora de ser explicados en la defensa oral. Williams y Power (2009) mostraron que en los textos escritos se tiende a aproximar cantidades numéricas al principio del documento, dando más precisión en las referencias posteriores a las mismas cantidades.

4. Como argumentó MacKay en su libro, "*la simplificación es una llave para el entendimiento, primero porque redondeando números son más fáciles de recordar y segundo porque con números redondeados se hacen cálculos más rápidos*" (MacKay, 2009). La simplificación brinda beneficios cognitivos haciendo que los números sean más fáciles de recordar y con razón. De ahí que la simplificación numérica pueda dar ventajas positivas tanto a personas con dificultades o sin ellas.

Vale la pena señalar que la simplificación de expresiones numéricas es una práctica normal en la edición de artículos periodísticos y una operación importante a la hora de realizar resúmenes de información. En efecto, no es raro ver titulares de noticias con expresiones numéricas vagas, como por ejemplo *"El fuego en Calcuta mató a docenas de personas"*, correspondiéndose con información mucho más precisa en el cuerpo de la noticia *"Al menos 89 personas han muerto en el fuego originado en la ciudad india de Calcuta..."*.

2.1.2. Procesos de lectura y razonamiento matemático

Leer es un proceso que consiste en convertir los símbolos gráficos en palabras, combinando las palabras en frases y finalmente combinando frases en un significado completo, sólo comenzando el siguiente nivel si el anterior ha sido completado. Este proceso va desde la acción más simple (la identificación de las formas) a la más compleja (la creación de significado global).

En la mayoría de las ocasiones, además de combinar el significante y el significado de las palabras o frases específicas, también se identifica la palabra en un contexto lingüístico más amplio. Es necesario contextualizar el significado, dado el número de significados que una palabra o frase pueden tener en función del contexto en el que aparece (Clemente y Domínguez, 2003).

Hay dos principios que explican el proceso de lectura de un texto (Solé, 1999):

- *El principio de inmediatez*: Comenzamos el proceso de lectura de un texto con información limitada e incompleta. A continuación creamos un significado tan rápido como nos sea posible, a pesar de que esto puede generar errores.

- *El principio de interactividad*: Los diferentes procesos de lectura son interactivos. Los procesos de decodificación y comprensión trabajan conjuntamente para crear una representación mental del significado. Leer implica utilizar el conocimiento semántico previo con el fin de predecir el texto y su significado antes de recibir las señales gráficas.

A continuación, enumeramos los factores que intervienen en los procesos de lectura (Ariles y Jiménez, 2011):

1. *Factores didácticos*: El proceso de lectura debe ser un método secuencial, desde la actitud con la que llevar a cabo la lectura hasta llegar a ser una forma de comunicación, con el dominio de estrategias básicas de comprensión lectora.

2. *Factores individuales*: Las características que diferencian a cada persona son cruciales para el proceso de lectura.

3. *Factores perceptibles*: La capacidad de percepción visual (discriminación figura-fondo, cierre visual, constancia visual) y auditiva (discriminación auditiva de fonemas y sílabas).

4. *Factores psicolingüísticos y metalingüísticos*: El proceso de adquisición y desarrollo del lenguaje y la capacidad de reflexionar sobre el propio uso del lenguaje.

5. *Factores cognitivos y metacognitivos*: Simbolización, análisis, síntesis, memoria y atención.

6. *Factores socio-emocionales*: Los hábitos de conducta, la voluntad de aprender, el reconocimiento de la utilidad de la lectura, la confianza y el autoestima.

Por lo tanto, el razonamiento matemático contribuye al desarrollo de la persona. Surge como una necesidad del hombre de comunicarse con los demás y expresar aspectos relacionados con el ambiente y sus necesidades: contar, medir y realizar operaciones matemáticas.

La mejor manera de entender el desarrollo de las habilidades matemáticas es recurrir a Piaget y Inhelder (1969), quienes distinguen las siguientes etapas en el desarrollo del pensamiento lógico:

- *Etapa de la inteligencia sensomotora (0-2 años)*: Supone la preparación funcional para el pensamiento lógico. En esta fase el niño investiga el entorno físico a través de los sentidos.

- *Etapa del pensamiento objetivo-simbólico (2-7 años)*: Comienza a aparecer la función de representación, con la capacidad de sustituir una acción por un símbolo. Durante los primeros siete años el niño descubre paulatinamente los principios de la invariación referidos a un objeto, al número, al espacio y al tiempo. Se clasifican los objetos según criterios como el color, la forma, el tamaño, etc.

- *Etapa del pensamiento lógico-concreto (7-12 años)*: Los procesos de razonamiento se vuelven lógicos y pueden aplicarse a problemas concretos o reales. En el aspecto social, el niño ahora se convierte en un ser verdaderamente social y en esta etapa aparecen los esquemas lógicos de la seriación, ordenamiento mental de conjuntos y clasificación de los conceptos de causalidad, espacio, tiempo y velocidad. Transcurridos los siete años, los niños son capaces de asumir y representar mentalmente las alteraciones de los números y de las cantidades, pero sólo de un modo reversible. Durante la primera infancia sólo son accesibles al niño los primeros números porque éstos corresponden a figuras perceptibles o manipulativas. Es después de los siete años cuando será capaz de acceder a la serie indefinida de números, y a las operaciones de suma, resta, multiplicación y división.

- *Etapa de operaciones formales (12-15 años)*: En esta etapa el niño ya es capaz de usar la lógica en la solución de los problemas que se le presentan. Logra la abstracción sobre conocimientos concretos observados que le permiten emplear el razonamiento.

El pensamiento lógico-matemático tiene una serie de determinantes que influyen directamente en el desarrollo normal de esta capacidad humana: noción de número, noción de espacio y tiempo, desarrollo del lenguaje y desarrollo de las funciones de atención y memoria (Piaget, 1921).

Para Piaget (1942), la construcción del concepto de número está relacionada con el desarrollo de la lógica, en cuanto a que en la génesis del número existe una organización mental previa al cálculo, sin la cual no sería posible adquirir los conocimientos básicos para construir el concepto. Así, la noción de número requiere la adquisición de las nociones de: conservación de la cantidad, reversibilidad de las operaciones, correspondencia término a término, seriación y clasificación e inclusión de la parte con el todo.

Para adquirir estas nociones el niño ha de dominar previamente una serie de conceptos básicos que influirán en las capacidades para relacionar números y cantidades mediante actividades de comparación u ordenación, así como unidades léxicas que distingan cantidades globales que se aproximen a la idea de número sin precisión. Encontrando entre ellas los cuantificadores (*pocos*, *muchos*, *casi*, *alguno*, etc.), identificadores (*diferente*, *igual*, *como*, etc.) y otros conceptos (*poner*, *juntar*, *vacío*, etc.).

La adquisición de conceptos matemáticos básicos y la realización de operaciones mentales de cálculo requieren también de un mínimo de atención y de memoria. Para facilitar el aprendizaje de los mismos, los contenidos matemáticos deben ser presentados de forma lúdica y sencilla suponiendo un estímulo para su atención, por lo que hay que explicitar su finalidad para la vida cotidiana.

2.1.2.1. Dislexia y discalculia

La *dislexia* es una dificultad neurológica caracterizada por problemas en la lectura a la hora de reconocer de manera fluida y exacta palabras, así como en la habilidad para deletrear y decodificar palabras. Esto imposibilita la correcta comprensión del texto que se lee (Vellutino et al., 2004). La personas con dislexia encuentran problemas para reconocer y recordar no solo letras sino también números (Newell y Booth, 1991; Cohen et al., 1994).

Una dificultad específica del aprendizaje que implica la dificultad innata para aprender o comprender la aritmética matemática es la *discalculia*. Es similar a la dislexia e incluye dificultades para comprender los números, aprender a manipular los números, aprender hechos matemáticos y una serie de otros síntomas relacionados como contar dinero, entender precios de los artículos, revisar un cambio recibido, retirar dinero en un cajero electrónico o recordar fechas (Butterworth, 2010).

Aunque la dislexia y la discalculia son dos dificultades distintas, ambas son comórbidas. La comorbilidad es un término médico, que indica que en este caso la discalculia existe simultáneamente pero independientemente con

otra condición médica, en este caso la dislexia. Las personas con dislexia son más propensas a tener dificultades de aprendizaje en el área de las matemáticas (Landerl et al., 2004).

2.1.3. Tareas principales en la simplificación de textos

En la simplificación de textos cabe destacar cuatro tareas principales sobre las que se está investigando a lo largo de estos años y que los trabajos desarrollados hasta ahora cubren de una manera u otra. Estas cuatro tareas son las siguientes:

1. Simplificación sintáctica: Transformar oraciones largas y complejas en oraciones simples e independientes, segmentando construcciones subordinadas y coordinadas, cambiando oraciones de pasiva a activa, etc.

2. Simplificación léxica: Reemplazar el vocabulario complejo, teniendo en cuenta el contexto, por palabras o expresiones más fáciles. Hay que considerar los casos de polisemia y resolver la ambigüedad. Se suelen usar bases de datos psicolingüísticas y diccionarios de sinónimos.

3. Eliminación de información: Prescindir de la información no necesaria para entender las ideas principales del texto. La información redundante se elimina para ayudar a la comprensión del texto.

4. Clarificación de información: Añadir explicaciones para los conceptos que se consideren más difíciles. Hay que decidir qué conceptos son difíciles pero importantes y por lo tanto no se deben eliminar, sino encontrar una definición o información necesaria para ayudar a su comprensión.

2.2. Prácticas existentes de simplificación manual de textos

Existen distintas iniciativas que han desarrollado procesos manuales de simplificación de textos siguiendo las pautas marcadas por las *Directrices para materiales en lectura fácil* de la IFLA (Freyhoff et al., 1998) y las pautas tituladas *El camino más fácil*, publicadas por la Asociación *Inclusion Europe* (Inclusion Europe Association, 1998). Todas ellas se mueven bajo el marco de la *Lectura Fácil*, un movimiento que promueve la creación de material (libros, documentos, páginas web, etc.) elaborados con especial cuidado tanto a nivel de contenido como de forma (formato, maquetación, márgenes, tipo de letra, espaciado, etc.) y que así las personas con dificultades lectoras puedan leerlos y entenderlos.

Estas directrices europeas están destinadas a autores, editores, responsables de información, traductores y otras personas interesadas en generar

información en lectura fácil. El acceso a la información es un aspecto fundamental para poder participar en la vida cotidiana. Sólo las personas bien informadas pueden influir o controlar las decisiones que afectan a sus vidas. Sin embargo, las actuales estructuras niegan el acceso a la información a un gran número de personas cuyas capacidades para la lectura, la escritura o el entendimiento están disminuidas por diversas razones. El objetivo de las directrices es que sirvan de estímulo para la generación de documentos en lectura fácil y así poder integrar en la sociedad de la información a toda la población europea.

Las características generales de los textos en lectura fácil son las siguientes:

- Utilizan un lenguaje simple y directo.
- Expresan una sola idea por frase.
- Evitan los tecnicismos, las abreviaturas y las iniciales.
- Estructuran el texto de manera clara y coherente.

Las directrices europeas recogen algunos de los pasos a seguir para elaborar documentos en lectura fácil. Se pueden dar dos situaciones distintas: la de disponer ya de un texto base que se quiere hacer accesible o la de generar un texto completamente nuevo. En ambos casos, hay que empezar pensando cuál es el grupo objetivo y la finalidad principal del texto que se intenta elaborar. A continuación recogemos los pasos indicados para el proceso de elaboración de un texto en lectura fácil:

1. Definir la finalidad de la publicación: Qué es lo que se quiere decir y por qué es importante para las personas del grupo objetivo.
2. Abordar el tema del contenido: Elaborar una lista con los aspectos clave de la publicación.
3. Elaborar el borrador del texto: Redactar el texto basándose en la lista de aspectos clave.
4. Comprobar que las personas del grupo objetivo entienden el borrador elaborado: Antes de generar la versión final del documento, una revisión con usuarios reales ayuda a corregir, mejorar y terminar de preparar la mejor versión posible.

Existen unas normas de tipo general que se deben observar a la hora de redactar un texto en lectura fácil:

- Usar un lenguaje sencillo y directo: Emplear las palabras más sencillas expresadas de la forma más simple.

- Evitar los conceptos abstractos: Usar ejemplos concretos que faciliten la comprensión del tema.
- Emplear palabras cortas relativas al lenguaje cotidiano hablado: evitar palabras largas difíciles de leer o pronunciar.
- Personificar el texto tanto como sea posible: Dirigirse a los lectores de manera directa y personal.
- Hacer uso de ejemplos prácticos: Pueden ser útiles para que las personas entiendan conceptos y relacionen información.
- Dirigirse a los lectores de manera respetuosa: Emplear lenguaje de adultos al escribir para personas adultas.
- Utilizar oraciones cortas en su mayoría.
- Incluir una sola idea principal en cada oración.
- Utilizar un lenguaje positivo: evitar negaciones y lenguaje negativo, ya que puede causar confusión.
- Emplear preferentemente la voz activa frente a la pasiva: El uso de voz activa hace que el documento sea más vivo y menos complicado.
- No dar por asumidos conocimientos previos sobre el tema en cuestión.
- Ser sistemático al utilizar las palabras: Utilizar la misma palabra para nombrar una misma cosa.
- Elegir signos de puntuación sencillos: Evitar el punto y coma, los guiones y las comas.
- No emplear el subjuntivo: El futuro incierto es impreciso y se presta a confusiones.
- Tener cuidado con el lenguaje metafórico si utiliza palabras de uso poco común.
- Tener cuidado con el uso de números: Las cifras largas o complicadas suelen ser incomprensibles. Para cifras pequeñas, utilizar siempre el número y no la palabra.
- No emplear palabras de otro idioma.
- Evitar el uso de referencias.
- Mencionar una dirección de contacto para obtener mayor información, cuando sea posible.

- Evitar el uso de jergas, abreviaturas e iniciales. Si es inevitable, explicar siempre su significado.

En los siguientes apartados presentamos las principales iniciativas de simplificación manual que han ido surgiendo desde el comienzo de la lectura fácil.

2.2.1. Lectura Fácil en los países nórdicos

En el año 1968 surgió en Suecia una iniciativa de adaptación de textos a lectura fácil cuyo resultado actual es la Fundación *Centrum för Lättläst* (Centro de Lectura Fácil)[8]. Ese mismo año publicaron su primer libro en lectura fácil en colaboración con la comisión de la Agencia Sueca de Educación. En el año 1984 lanzaron el primer periódico en lectura fácil de forma experimental, titulado *8Sidor* (Ocho páginas), que empezó a publicarse de forma permanente desde 1987. En 1991 crearon su propia editorial especializada para este tipo de publicaciones. Hasta el año 1994 se habían publicado unas 330 obras en lectura fácil y la producción media era de entre 15 y 20 nuevas publicaciones al año. En 1997 se transforma la Fundación en el *Centro de Lectura Fácil* por mandato parlamentario y se crea un departamento de adaptación de textos administrativos. La financiación del centro depende de los ingresos de las publicaciones editadas y los fondos de subvenciones estatales.

Su director, Bror Tronbacke, fue el redactor de las *Directrices para materiales de lectura fácil*, publicadas en 1997 por la IFLA. El centro sueco es, posiblemente, el más antiguo y mejor organizado del mundo. Su experiencia se ha extendido de forma similar en los países vecinos, Noruega y Finlandia.

En Noruega, la iniciativa se llama *Leser søker bok* (Lector busca libro) y es una alianza de 20 organizaciones, que incluyen editoriales y asociaciones de personas con discapacidad. Fue creada en 2003 y ha editado unos 60 títulos.

Por su parte, en Finlandia existen dos centros, uno de lengua finlandesa y otro de lengua sueca, lenguas co-oficiales en el país. Ambos centros están vinculados a las organizaciones de personas con discapacidad intelectual. Publican libros, seminarios y folletos de información de interés ciudadano en lectura fácil.

Como fruto de la experiencia escandinava surgió en 2004 la *International Easy-to-read Network*[9] (Red Internacional de lectura fácil), que tiene como impulsoras a las organizaciones finlandesa, noruega y sueca. En la actualidad, cuenta con más de 60 organizaciones y particulares asociados de 30 países de todo el mundo.

[8] http://www.lattlast.se/
[9] http://wordpress.easytoread-network.org/

Figura 2.1: Logotipo europeo de lectura fácil diseñado por *Inclusion Europe*

2.2.2. *Inclusion Europe*: el marco europeo de personas con discapacidad intelectual

La organización *Inclusion Europe*[10] se creó en 1988, tiene la sede en Bruselas y es el punto de encuentro de las asociaciones de personas con discapacidad intelectual en la Unión Europea. Agrupa a las organizaciones de 40 países europeos e Israel. Su objetivo es luchar por la igualdad de derechos y la plena inclusión de personas con discapacidad intelectual y sus familias en todos los aspectos de su vida.

En 1998 elaboró la guía *El camino más fácil: Directrices europeas para generar información de fácil lectura destinada a personas con discapacidad intelectual*, en la que propone las pautas para desarrollar un proyecto de redacción original en lectura fácil o la adaptación de textos a esta técnica. Además, diseñó un logotipo europeo de lectura fácil (Figura 2.1) para identificar todos los textos redactados que siguieran sus pautas. Trabajan redactando y adaptando textos a lectura fácil en 20 lenguas europeas. Publican cada día una revista online *e-Include*[11] que ofrece noticias, eventos y artículos sobre diferentes temas relacionados con la discapacidad intelectual.

2.2.3. Proyecto *Pathways*

El proyecto Pathways I[12] (2007- 2009) tiene como finalidad la necesidad de formalizar la lectura fácil como una herramienta de inclusión de las personas con discapacidad. Promovido por *Inclusion Europe* junto con sus socios de Austria, Alemania, Finlandia, Irlanda, Lituania, Portugal y Escocia, intentaron abordar la lectura fácil de forma global, no sólo atendiendo al método de redacción y evaluación, sino también pensando en las personas con discapacidad intelectual como agentes que redactan textos y en los profesores que participan en programas de formación continua. La idea continuó con el proyecto Pathways II[13] (2011- 2013) ampliando sus materiales a otros países europeos como Croacia, República Checa, Estonia, Hungría, Italia,

[10] http://inclusion-europe.org/es
[11] www.e-include.eu
[12] http://inclusion-europe.org/en/projects/past-projects/pathways-i
[13] http://inclusion-europe.org/es/proyectos/pathways-ii

Eslovenia, Eslovaquia y España.

2.2.4. Asociación Lectura Fácil de Barcelona

La Asociación de Lectura Fácil[14] con sede en Barcelona fue la primera de estas características en crearse en España. Es una entidad sin ánimo de lucro, que trabaja para acercar la lectura a las personas con dificultades lectoras. Se creó en 2002 y cuenta con más de 1500 suscriptores, más de 122 libros de lectura fácil y unos 90 clubes de lectura fácil para promover esta actividad entre grupos con dificultades lectoras.

Desde 2005 es miembro de la International Easy-to-Read Network. Parte del principio de democracia lectora: todo el mundo debe tener acceso a la información, a la literatura y a la cultura para poder participar de forma activa y responsable en la sociedad. La Asociación ha asesorado a entidades públicas y asociaciones para adaptar folletos y textos informativos a lectura fácil.

2.2.5. Portal web *Noticias fácil*

El portal web *Noticias fácil*[15] publica noticias, libros y documentos en lectura fácil para acercar la información a todas las personas. Está hecho por la Fundación ONCE[16] y se dirige a personas con discapacidad intelectual o cognitiva y a personas con problemas de comprensión lectora. Conocer lo que pasa es muy importante, pero hay personas que no entienden las noticias que están escritas en los periódicos porque tienen un lenguaje complicado.

En este portal, las noticias son cortas, no tienen palabras complicadas y todo el mundo puede leerlas sin cansarse. Los principales objetivos son que la información y las noticias diarias no tengan barreras, que puedan llegar a todo el mundo y que las personas con discapacidad intelectual o cognitiva puedan participar, opinar y crear sus propias noticias. Pretende además ser un punto de encuentro y opinión a través de encuestas y blogs, así como ayudar en la mejora de las habilidades de comprensión y comunicación de personas con discapacidad intelectual.

2.2.6. FEAPS

La Federación de organizaciones en favor de personas con discapacidad intelectual (FEAPS)[17] tiene como misión contribuir, desde su compromiso ético, con apoyos y oportunidades, a que cada persona con discapacidad intelectual o para el desarrollo y su familia puedan desarrollar su proyecto de

[14] http://lecturafacil.net
[15] http://www.noticiasfacil.es
[16] http://www.fundacionONCE.es
[17] http://www.feaps.org

calidad de vida, así como promover su inclusión como ciudadano de pleno derecho en una sociedad justa y solidaria. Es una entidad sin ánimo de lucro que tiene su sede en Madrid desde 1978 y cuya acción se traduce en proveer servicios, defender derechos y ser agente de cambio social. FEAPS cuenta con 884 entidades, con 17 federaciones autonómicas, 235.000 familiares, 139.000 personas con discapacidad intelectual o del desarrollo, 4.000 centros y servicios, 40.000 profesionales y 8.000 personas voluntarias.

Hasta la fecha han realizado varias publicaciones en lectura fácil, aplicando la metodología definida por la IFLA. Además, el departamento de comunicación complementa las notas de prensa que lanza a los medios con su versión en lectura fácil. Por otra parte hay que señalar que los educadores y psicólogos de FEAPS que trabajan a diario con personas con discapacidad intelectual han encontrado un gran apoyo en la lectura fácil.

2.3. Aproximaciones a la simplificación automática de textos

En todas las iniciativas presentadas en la sección anterior se realiza la adaptación de textos de manera manual para generar las versiones en lectura fácil para que la información sea accesible. Pero la tarea de simplificación manual es un trabajo muy costoso en tiempo y recursos. Hoy en día la información se genera muy rápidamente y es imposible una adaptación manual accesible en tiempo real. Con el objetivo de solventar este problema surge la simplificación automática de textos.

En esta sección se describen los principales sistemas de simplificación de textos en orden cronológico aproximado, destacando su novedad y discutiendo cómo el campo ha evolucionado con el tiempo. En los recientes años ha crecido la idea de aplicar traducción automática en el proceso de simplificación de textos, considerada como una traducción monolingüe, ya que es un único idioma, traduciéndose de la versión original a la versión simplificada. Impulsada por la nueva disponibilidad de corpus de textos simplificados, ha surgido una dicotomía entre sistemas diseñados de forma manual con reglas escritas a mano y los enfoques que aprenden a partir de corpus utilizando modelos estadísticos. Todos ellos han explorado una gran variedad de representaciones lingüísticas para codificar las operaciones de simplificación, ya sean a nivel sintáctico o léxico.

Los primeros estudios de la simplificación automática de textos cubren mucho terreno, la exploración de los sistemas, los sistemas que aprenden reglas de simplificación del texto hecho a mano (y de hecho, adoptan ideas de traducción automática) y el análisis de las cuestiones de simplificación léxica y sintáctica, así como la coherencia del texto. Algunas de las ideas de estos trabajos han sido redescubiertas en los últimos años, mientras que otras han sido olvidadas.

Uno de los primeros trabajos que tuvieron como objetivo de investigación la simplificación de textos en inglés, fue el trabajo de Chandrasekar et al. (1996). Su principal motivación era reducir la longitud de la oración en la fase del preprocesamiento del texto a la hora de analizarlo. Definieron el proceso de simplificación en dos fases: análisis y transformación. En la primera fase se obtiene la representación estructural de la oración y en una segunda fase se aplica una secuencia de reglas para identificar y extraer los componentes que pueden ser simplificados. En una primera aproximación definieron las reglas de transformación sintáctica manualmente, y en una segunda aproximación las aprendieron de un corpus alineado que crearon con las versiones originales y su correspondiente versión simplificada. La idea que se persiguió en el trabajo era que si un texto era complejo podía convertirse en un texto más simple aplicando un proceso de simplificación a nivel de oración. El proceso consistía en identificar componentes de una oración que podían ser tratados por separado y transformarlos en otros más simples. Se asumía que en el proceso de simplificación se producía una pérdida de información con respecto al texto original.

El proyecto PSET (*Practical Simplification of English Text*)(Carroll et al., 1998) fue quizá el primero en aplicar tecnologías de lenguaje natural para personas con dificultades lectoras. Su objetivo era simplificar las noticias de prensa en inglés para personas con afasia. Estaba formado por tres componentes: uno de simplificación sintáctica, uno de resolución de anáforas y otro de simplificación léxica. Para las transformaciones sintácticas usaba reglas definidas manualmente sobre los árboles sintácticos del parser: convertía oraciones pasivas en activas, dividía oraciones coordinadas, eliminaba las oraciones relativas y en general sustituía las oraciones largas por dos o más oraciones cortas. Para llevar a cabo la simplificación léxica usaba la base de datos léxica *WordNet* (Miller et al., 1990). Para cada palabra creaba un archivo con los sinónimos de la palabra y elegía la palabra más apropiada, la de mayor frecuencia, usando la base de datos psicolingüística desarrollada en Oxford, *The Oxford Psycholinguistic Database* (Quinlan, 1992).

Usando reglas basadas en patrones de simplificación, en el trabajo de Canning (2000) se presentó el sistema SYSTAR (SYntactic Simplification of Text for Aphasic Readers) perteneciente al proyecto PSET. Este módulo era el encargado de separar las oraciones, activar las oraciones que estaban en pasiva y resolver y reemplazar los pronombres anafóricos que ocurrían con frecuencia. Para cada oración se realizaba un proceso recursivo de aplicación de cada regla hasta que no había coincidencia con todas las reglas que habían sido aplicadas.

La tesis doctoral de Dras (1999) fue otro trabajo importante en el campo de la simplificación. Entre sus principales contribuciones se encuentra una lista de operaciones de paráfrasis para el inglés. Utilizó el formalismo de *Tree Adjoining Grammar* (TAG) para representar una oración y fue el primero

en usar *Integer Programming* para generar un texto que cumpliera unas restricciones externas impuestas. Estas dos ideas han sido redescubiertas en los recientes trabajos de simplificación de texto ((De Belder et al., 2010), (Woodsend y Lapata, 2011), (Siddharthan y Angrosh, 2014)).

En el trabajo de tesis doctoral de Siddharthan (2003) se presentó el proceso de simplificación sintáctica automática para reducir la complejidad de un texto en inglés. Describió cómo la simplificación sintáctica se consigue a partir de un análisis de un conjunto de reglas creadas manualmente y de un análisis detallado a nivel de discurso para poder reescribir el texto. El trabajo consideró el tratamiento de oraciones relativas, de aposición, coordinadas y subordinadas. Además en su trabajo señaló la necesidad de un componente de regeneración en el proceso de simplificación de textos para mostrar cómo ciertas reestructuraciones sintácticas de un texto pueden significar alteraciones a nivel de estructura discursiva del texto. Formalizó las interacciones entre sintaxis y discurso durante el proceso de simplificación de un texto y mostró cómo conservar la cohesión y la coherencia en un texto.

Inui et al. (2003) propusieron un sistema basado en reglas para la simplificación de textos en inglés dirigida a personas sordas. El objetivo de este sistema era aplicar transformaciones sintácticas y léxicas a nivel de paráfrasis a un texto dado para generar un texto más fácil de entender para las personas sordas. Este tipo de personas en particular tienen dificultades en la comprensión debido a que su lengua materna, el lenguaje de signos, es esencialmente visual.

El problema de alineación de oraciones en un corpus monolingüe fue abordado en el trabajo de Barzilay y Elhadad (2003). A partir de la alineación automática de un corpus se proporcionó un recurso valioso para el aprendizaje de reglas de reescritura. Además, se incorporó el contexto en la búsqueda de una alineación óptima, obteniendo muy buenos resultados en los experimentos llevados a cabo.

Daelemans et al. (2004) aplicaron simplificación automática a nivel de oración para generar subtítulos de programas de televisión en holandés y en inglés para ayudar a espectadores sordos. Compararon dos métodos de simplificación, uno basado en el aprendizaje a partir de un corpus paralelo y otro basado en reglas definidas manualmente.

Williams y Reiter (2005) presentaron un sistema de generación de texto que adaptaba su salida para lectores con baja alfabetización. Definieron reglas de restricciones para evitar combinaciones ilegales y reglas de optimización que expresaban las preferencias de legibilidad. Vieron que las decisiones basadas en el conocimiento de microplanificación mejoraban la legibilidad del texto para este colectivo concreto.

Elhadad (2006) utilizó el corpus de frecuencia de *Reuters Health E-line news-feed*[18], un recurso con el que los periodistas resumen publicaciones téc-

[18] www.reutershealth.com

nicas tales como ensayos clínicos para los lectores novatos, para determinar los términos médicos difíciles para este tipo de lectores.

En el trabajo de Petersen y Ostendorf (2007) se llevó a cabo un análisis de un corpus paralelo de artículos de noticias para aprender qué tipo de cambios realiza la gente cuando simplifica textos para personas que están aprendiendo una lengua. El corpus está formado por los artículos originales junto con sus correspondientes versiones abreviadas desarrolladas por *Literacyworks*[19] como parte de una página web de alfabetización para hablantes nativos que tienen pocas habilidades lectoras. Para entender las técnicas que los autores usan cuando editan cada artículo original para crear la versión abreviada, se realizó una alineación manual de las oraciones para cada par de artículos. Casi todas las oraciones originales estaban alineadas con una o más oraciones en la versión simplificada, aunque algunas oraciones originales eran eliminadas y no aparecían en la versión simplificada. Se realizó un análisis comparando los artículos original y abreviado, mostrando la importancia de las características sintácticas, además de la longitud de las oraciones, para decidir si separar oraciones, y la posición y la información redundante, para decidir si la oración original se mantiene o se elimina en la versión abreviada.

Los trabajos de Aluísio et al. (2008) y Candido et al. (2009) presentaron el sistema *PorSimples* para el idioma portugués, desarrollado para ayudar a lectores con baja alfabetización a procesar documentos de la web. En el proyecto usaron diferentes técnicas de adaptación de texto: resumen automático para hacer los textos más cortos, simplificación léxica para reemplazar las palabras complejas por otras más simples, simplificación sintáctica para separar oraciones complejas y elaboración del texto para añadir información de apoyo. Propusieron un conjunto de operaciones para simplificar 22 construcciones sintácticas a partir de un análisis manual de textos simplificados.

A partir del desarrollo de las *Directrices para materiales en lectura fácil* de la IFLA (Freyhoff et al., 1998), en el trabajo de Bautista et al. (2009) se utilizaron un subconjunto de estas pautas para diseñar e implementar reglas automáticas a nivel de transformaciones sintácticas y substituciones léxicas. Usando las métricas de legibilidad de los textos en inglés se medía la complejidad del texto antes y después de la simplificación y se veía la mejora en las versiones simplificadas.

El trabajo de Zhu y Gurevych (2010) examinaba la *Wikipedia*[20] en inglés y su versión simplificada *Simple English Wikipedia*[21] como una aproximación basada en datos para la tarea de simplificación de textos. Propusieron una solución probabilística basada en sintaxis para compararla con una solución de referencia no simplificada y una solución basada en traducción automática basada en oraciones.

[19] http://www.literacyworks.org/
[20] http://en.wikipedia.org
[21] http://simple.wikipedia.org

Specia (2010) fue la primera en aplicar *Phrase Based Machine Translation (PBMT)* a la tarea de simplificación de textos, en su caso para el portugués. Consta de un proceso en dos etapas. En la primera etapa se realiza la alineación a nivel de palabra. La segunda etapa se centra en la descodificación para encontrar la mejor traducción de la oración inicial a la oración objetivo.

En el trabajo de Yatskar et al. (2010) se llevó a cabo también una revisión de las simplificaciones léxicas realizadas en la *Simple English Wikipedia* con el objetivo de aprender de dichas transformaciones.

De Belder et al. (2010) usaron un sistema basado en reglas para simplificar las construcciones sintácticas de aposición, cláusulas relativas, subordinación y coordinación. Representaron las frases utilizando los árboles sintácticos proporcionados por el Stanford Parser (Klein y Manning, 2003). Siguieron la propuesta de Dras (1999) para decidir qué frases simplificar a través de la satisfacción de restricciones a nivel de todo el documento, en lugar de a nivel de frase.

En el trabajo de Kandula et al. (2010) se identificaron los términos difíciles en el texto y se simplificaron reemplazándolos por sinónimos más fáciles o usando una explicación con términos relacionados utilizando una frase corta para describir la relación entre el término difícil y el término seleccionado.

El proyecto *Simplext*[22] (Sistema automático de transformación de contenidos en textos de fácil lectura) (Saggion et al., 2011) tenía como objetivo principal desarrollar un producto de apoyo para la simplificación de textos en español para colectivos de personas que tienen necesidades especiales de lectura y comprensión. A partir de una metodología de simplificación manual definida por Anula (2007, 2008) se consiguió reducir la complejidad del texto. Se consideraron dos tipos de operaciones de simplificación: a nivel de estructuras sintácticas y a nivel de simplificaciones léxicas, a partir de las cuales se definieron reglas que se aplicaban automáticamente en el proceso de simplificación del texto original.

El problema de simplificación fue abordado como un problema de traducción automática de inglés a inglés en el trabajo de Coster y Kauchak (2011) siguiendo la metodología de *Phrase Based Machine Translation (PBMT)* con una etapa de descodificación diferente, permitiendo alinear frases originales con frases objetivos vacías, debido a la operación de eliminación de información. Utilizaron un corpus de oraciones alineadas extraído de alinear la *Wikipedia* en inglés y la versión simplificada de la misma. Este conjunto de datos contenía las operaciones de transformación incluyendo reordenación, uso de otras palabras, inserción o eliminación de información. Introdujeron un nuevo modelo de traducción para simplificación de textos que extiende la aproximación de traducción automática basada en frases que incluía la operación de eliminación. El principal objetivo era, dada una oración, producir una oración simplificada con un vocabulario y estructura simple preservando

[22] http://www.simplext.es/

el significado y las principales ideas de la oración original.

En el trabajo de Bautista et al. (2011c) se presentó un análisis de un corpus paralelo que contiene versiones de textos originales y sus correspondientes versiones simplificadas manualmente. Se utilizó el corpus creado por Barzilay y Elhadad (2003) y el objetivo fue identificar qué tipo de transformaciones son usadas para crear las versiones simplificadas, para su futura automatización a partir del diseño e implementación de un conjunto de reglas que permita realizar dichas transformaciones.

El trabajo de Walker et al. (2011) se centró en la simplificación léxica. Señalaron la ambigüedad como otro factor a tener en cuenta en el proceso de simplificación de texto. Se dieron cuenta de que había una correlación entre la frecuencia de las palabras del corpus y el número significados que tenían en WordNet. Vieron que los lectores preferían palabras no ambiguas pero menos frecuentes frente a palabras más comunes pero ambiguas.

Biran et al. (2011) definieron la complejidad de una palabra del corpus como la proporción de su frecuencia en la *Wikipedia* en inglés y su correspondiente versión simplificada (*Simple English Wikipedia*). Para calcular su dificultad multiplicaban este valor por la longitud de la palabra. Demostraron que este método mejoraba la propuesta de reemplazar las palabras con su sinónimo más frecuente calculado por WordNet, mejorando así la gramaticalidad de la salida, la preservación del significado y la simplicidad.

Woodsend y Lapata (2011) presentaron un modelo basado en gramáticas causi-síncronas y programación lineal entera. Con las gramáticas generaban todas las posibles operaciones de reescritura para un árbol sintáctico y con la programación lineal entera, usando restricciones, seleccionaban la simplificación más apropiada.

Siguiendo la idea presentada en el trabajo de Coster y Kauchak (2011), Wubben et al. (2012) extendió el segundo paso en *Phrase Based Machine Translation (PBMT)* con un estado de descodificación diferente. El objetivo era encontrar alineaciones de frases donde la frase simple es lo más diferente posible a la frase original, con la intuición de que tales paráfrasis tenían más probabilidades de simplificar el texto. Nótese que *PMBT* puede sólo llevar a cabo un pequeño conjunto de operaciones, como la sustitución léxica, la eliminación y simples paráfrasis. Para las operaciones de reordenación y división de oraciones no es muy adecuada.

Hay sistemas que usan árboles de dependencias para representar las oraciones y definen sobre ellos reglas de transformación. Es el caso del sistema presentado por Bott et al. (2012) para realizar simplificación en español, que permite simplificar cláusulas relativas, construcciones coordinadas y de participio.

El análisis de oraciones largas es la raíz de los problemas en las aplicaciones de traducción automática. Con el objetivo de resolver estos problemas se aplican simplificaciones sintácticas. En el trabajo de Aranzabe et al. (2012)

propusieron la primera simplificación automática para el euskera usando reglas específicas para simplificar las estructuras sintácticas de ese lenguaje.

Para el caso del idioma francés cabe destacar el trabajo de Seretan (2012) en el que se centraron en reducir la complejidad sintáctica, y el trabajo de François y Fairon (2012) donde presentaron una nueva fórmula para medir la legibilidad de un texto en francés.

Barbu et al. (2013) presentaron el proyecto *FIRST (Flexible Interactive Reading Support Tool)* donde se desarrolló una herramienta para asistir a personas autistas para adaptar los documentos escritos en un formato que sea más fácil de leer y entender para ellos. La herramienta aplica una serie de transformaciones automáticas para identificar y eliminar los obstáculos que les producen problemas en la lectura y comprensión de textos.

Saquete et al. (2013) desarrollaron un proyecto centrado en el tratamiento de textos educativos en español con la finalidad de reducir las barreras lingüísticas que dificultan la comprensión lectora a personas con deficiencias auditivas, o incluso a personas que están aprendiendo una lengua distinta a su lengua materna.

Para el colectivo de personas sordas, distintos trabajos han sido presentados en diferentes idiomas. El trabajo de Lozanova et al. (2013) propone un sistema basado en reglas para la simplificación automática de textos en búlgaro. Para el coreano, el trabajo de Chung y Park (2013) convierte oraciones complejas en otras más simples y muestra las relaciones con una representación gráfica.

Recientemente, las últimas propuestas que abordan la tarea de simplificación de textos, como el trabajo de Siddharthan y Angrosh (2014), retoman la idea de usar gramáticas de dependencias síncronas combinándolas con gramáticas construidas manualmente, para reglas sintácticas, y gramáticas adquiridas automáticamente, para reglas sintácticas y paráfrasis. Además, se sigue trabajando en la línea de usar reglas definidas manualmente, como en el trabajo de Brouwers et al. (2014) donde usan reglas a base de una tipología de reglas de simplificación extraídas manualmente de un corpus de textos simplificados en francés. Evans et al. (2014) presentan la evaluación de reglas de simplificación sintáctica para personas con autismo para reescribir las oraciones complejas. En el trabajo de Siddharthan (2014) se revisa la disciplina de simplificación de textos presentando un estudio de los distintos sistemas implementados hasta ahora.

Podemos observar que en todos los sistemas de simplificación automática desarrollados hasta ahora tienen un papel fundamental, de una manera u otra, el idioma con el que trabajan, como el usuario final al que va dirigida la simplificación, el tipo de texto y el nivel de dificultad al que se adaptan los textos. Cada sistema considera un conjunto de operaciones de simplificación a distintos niveles, sintácticos o léxicos, para llevar a cabo la adaptación del texto original. En capítulos posteriores podremos ver cómo estas variables

son consideradas en la propuesta de trabajo que presentamos en esta tesis.

2.3.1. Trabajos centrados en la simplificación de información numérica

Dentro de los trabajos de simplificación de textos prestamos especial atención a los que se han centrado en el tratamiento de la información numérica, ya que el trabajo que se presenta en esta tesis se engloba dentro de la simplificación de expresiones numéricas. A continuación presentamos los trabajos más relevantes en el área de investigación del procesamiento de la información numérica.

Bisantz et al. (2005) realizaron un estudio para analizar la representación de la información probabilística. Manejaban dos variables, una en relación al formato de la información (borrosa, en iconos, en frases lingüísticas, en expresiones numéricas) y la otra en relación al nivel de especificación (en la que el número y tamaño de los pasos discretos en la que la información probabilística fue asignada). La representación lingüística de la incertidumbre (como *raramente*, *probablemente*) ha sido representada de forma vaga (gráfica o lingüística) en comparación con la representación numérica precisa de la probabilidad (expresiones numéricas, expresiones gráficas anotadas con formatos numéricos). Se comparó el uso de diversas expresiones lingüísticas y numéricas para valores de probabilidad y apenas se encontraron pequeñas diferencias según la opinión de los participantes. Investigaciones sobre probabilidad lingüística (Budescu y Wallsten, 1995) tienen como hipótesis de trabajo que para usar representaciones lingüísticas a la hora de tomar decisiones, las personas convierten estas representaciones en estimaciones numéricas con valores concretos.

En el trabajo de Peters et al. (2007) examinaron el concepto de *numeracy* (habilidad para la aritmética), por qué es importante esta habilidad para las decisiones de atención a la salud y cuáles son las mejores prácticas para la presentación de la información numérica en este contexto. Para ello investigaron acerca de la influencia de la información numérica en la comprensión y de qué estrategias existen para presentar la información numérica al paciente.

El tratamiento de la información numérica en el área de la predicción del tiempo atmosférico fue recogido en el trabajo de Dieckmann et al. (2009). Se centraron en los marcadores de decisión que a menudo se presentan con las evaluaciones de probabilidad (por ejemplo, *hay un 15 % de posibilidades de que un cambio atmosférico ocurra en los próximos tres meses*) y con el apoyo de la narrativa en el dominio atmosférico y del tratamiento de la información numérica para dar un diagnóstico, preciso y fiable. Realizaron un par de estudios para explorar cómo los marcadores de decisión varían en narrativa e información numérica a la hora de realizar un pronóstico.

El proyecto *NumGen*[23] (*Generating Intelligent Descriptions of Numerical Quantities for People with Different Levels of Numeracy*) (Williams y Power, 2009, 2010) tuvo como finalidad determinar cómo presentar la misma información numérica de diferentes formas para diferentes usuarios. Para ello desarrollaron un sistema en *Prolog* basado en restricciones que dada una proporción de entrada genera un conjunto de posibles versiones equivalentes en distintas representaciones matemáticas. Además, como parte del proyecto, construyeron un corpus de artículos de prensa que tenían un alto contenido de expresiones numéricas. Destacamos este proyecto para el inglés, debido a que sistemas previos habían trabajado sólo en la variación de la representación de los datos numéricos limitándose a elegir entre dígitos o letras, como es el caso de los sistemas *SkillSum* y *GIRL* (Williams y Reiter, 2008). En cambio otros sistemas generaban descripciones numéricas para algunos grupos de usuarios pero no podían variarlas para otros. Por ejemplo, *SumTime* (Reiter et al., 2005) describe datos numéricos para profesionales del tiempo atmosférico pero no genera descripciones numéricas comprensibles para personas no profesionales.

Estudios previos han demostrado que las personas eligen información precisa frente a información difusa, porque les da sentido de seguridad y hacen que su ambiente sea más predecible. Sin embargo, en el trabajo de Mishra et al. (2011) mostraron que los entornos borrosos de información vaga (intervalos) pueden ayudar a los individuos a realizar mejores comparaciones de información que si la información se da de forma precisa. Actualmente vivimos rodeados de dispositivos que nos permiten acceder a información precisa en cada momento, podemos saber cuántos kilómetros hemos recorrido, cuántas calorías tiene lo que hemos comido, a qué distancia se encuentra un punto determinado, lo que nos proporciona un nivel de exactitud que nos da seguridad. Pero en los experimentos realizados en el trabajo encuentran que la información difusa en muchas ocasiones nos sirve de mejor manera que la información precisa. Esto es debido a que la información difusa da a los individuos libertad y flexibilidad a la hora de percibir la información y formar así expectativas de acuerdo con sus deseos.

Durante el proyecto *Simplext*[24] se desarrolló un estudio de la simplificación manual de textos en español (Saggion et al., 2011). Se definió una arquitectura computacional para la simplificación automática de textos, donde se especificaron y se implementaron un conjunto de técnicas de procesamiento de lenguaje natural. Para ello se determinaron los aspectos lingüísticos de los textos escritos susceptibles de simplificación formal orientada a la mejora de la legibilidad y comprensibilidad en español escrito. Se creó un corpus paralelo de textos informativos, original y simplificado, alineados a nivel de oración (Bott y Saggion, 2011a,b). En colaboración con ese proyecto, como

[23] http://mcs.open.ac.uk/sw6629/numgen/
[24] http://www.simplext.es/

parte del trabajo de esta tesis, centrándonos en el tratamiento de la información numérica, se desarrolló un componente basado en reglas para reescribir las expresiones numéricas que aparecen en los textos. Se realizó un estudio para identificar las estrategias de simplificación usadas para simplificar expresiones numéricas, a partir del corpus paralelo y de un estudio que se realizó con expertos (Bautista y Saggion, 2014b).

Resumen y conclusiones

En este capítulo se ha presentado la tarea de simplificar un texto a partir de las principales operaciones identificadas en el proceso de simplificación. Se ha mostrado la necesidad de esta tarea y, en particular, la necesidad de simplificar expresiones numéricas, que es donde se enmarca el trabajo de esta tesis.

Además, se han revisado las distintas propuestas de simplificación manual desarrolladas hasta ahora por distintas iniciativas nacionales y europeas. También se han revisado las distintas aproximaciones de simplificación automática de textos implementadas en el área, haciendo especial hincapié en aquellas que tratan información numérica.

Con todo esto, nuestra propuesta de trabajo se enmarca en la simplificación automática de expresiones numéricas, implementando un modelo computacional a partir de un modelo genérico de proceso (presentado en el capítulo 4) que hemos definido con lo aprendido de los modelos existentes.

En el siguiente capítulo se presentan las herramientas y recursos necesarios para el análisis de los textos, junto con las herramientas específicas para el tratamiento de expresiones numéricas.

Capítulo 3

Herramientas y recursos

Distintas herramientas y recursos entran en juego a la hora de llevar a cabo la tarea de simplificar un texto. En este capítulo revisamos los corpus como recurso de simplificación, presentamos diversas herramientas de análisis de texto y nos centramos en las herramientas específicas para el tratamiento de expresiones numéricas usadas en el trabajo de esta tesis.

3.1. Corpus como recurso de simplificación

La tarea de simplificación de textos se puede considerar como un problema de traducción entre dos subconjuntos de textos: el original y el simplificado. Tras la idea de generar una versión simplificada del texto original se persigue transmitir la misma información pero de una manera más simple, de igual manera que se hace cuando se quiere transmitir en otro lenguaje en un proceso de traducción.

Para llevar a cabo esta idea de tarea de traducción es importante contar con un conjunto de textos originales que nos permitan aplicar las transformaciones deseadas para generar la versión simplificada. En el campo del procesamiento del lenguaje natural a este conjunto de textos se le llama *corpus*. Los textos están agrupados de acuerdo con su contenido, ya que el contexto en el que se trabaja y el conjunto de textos con el que se cuenta es muy importante.

Un corpus importante dentro del área de la simplificación de textos es el creado por Barzilay y Elhadad (2003). Es un corpus paralelo monolingüe en inglés, cuyas fuentes de información son la Enciclopedia Británica y la Enciclopedia Elementaria[1]. Los textos están alineados a nivel de oración, donde incorporaron una descripción básica del contexto para encontrar la manera óptima de realizar la alineación. Cada par de textos describen una misma ciudad, pero los que proceden de la Enciclopedia Británica son artículos más

[1] http://www.britannica.com/

detallados para adultos, mientras que los textos de la Enciclopedia Elementaria corresponden a una versión adaptada para niños. Hay un total de 2.600 artículos en la versión Elementaria diseñados para ayudar a los estudiantes de 6 a 10 años.

Para el trabajo presentado en esta tesis era necesario contar con un corpus rico en expresiones numéricas en inglés para poder validar nuestras hipótesis de trabajo. En nuestro caso utilizamos el corpus creado dentro del proyecto *NumGen*[2] ya presentado en la sección 2.3.1. El corpus usado en este proyecto es un conjunto de textos correspondientes a noticias de prensa y artículos científicos que presentan la misma información numérica en diversas formas matemáticas y lingüísticas, y que incluyen ejemplos de cardinales, ordinales, fechas, decimales, fracciones, porcentajes y ratios. El corpus está formado por 110 artículos, con 2.648 oraciones, 54.584 palabras, 1.888 expresiones numéricas y 404 modificadores numéricos.

Como parte del proyecto *Simplext*[3], presentado en la sección 2.3.1, orientado hacia el desarrollo de un sistema de simplificación automática de textos en español para los lectores con discapacidad cognitiva, se recopiló un corpus que consiste en 110 textos informativos, en el dominio de noticias internacionales y de cultura, cedidos por la agencia española de noticias Servimedia[4]. La metodología adoptada en el proyecto se basó en la creación del corpus y sus simplificaciones manuales con el fin de realizar un estudio que permitiera discernir qué manipulaciones serían necesarias para obtener una simplificación automática apropiada. Este corpus es rico en expresiones numéricas y es utilizado en el trabajo presentado en esta tesis para validar nuestras hipótesis de trabajo en español.

3.2. Herramientas de análisis de texto

A la hora de llevar a cabo el análisis de un texto se utilizan diferentes herramientas para cubrir las distintas tareas que hay que realizar. En esta sección presentamos los analizadores sintácticos más destacados para el inglés y para el español. Los analizadores sintácticos son los encargados de convertir el texto de entrada en otras estructuras, comúnmente árboles, que son más útiles para el posterior análisis. Además, presentamos GATE (Cunningham et al., 2002), que es un conjunto de herramientas de procesamiento de lenguaje natural en una plataforma desarrollada en Java que se usa para muchas tareas de computación relacionadas con el lenguaje.

[2] http://mcs.open.ac.uk/sw6629/numgen/
[3] www.simplext.es
[4] http://www.servimedia.es/

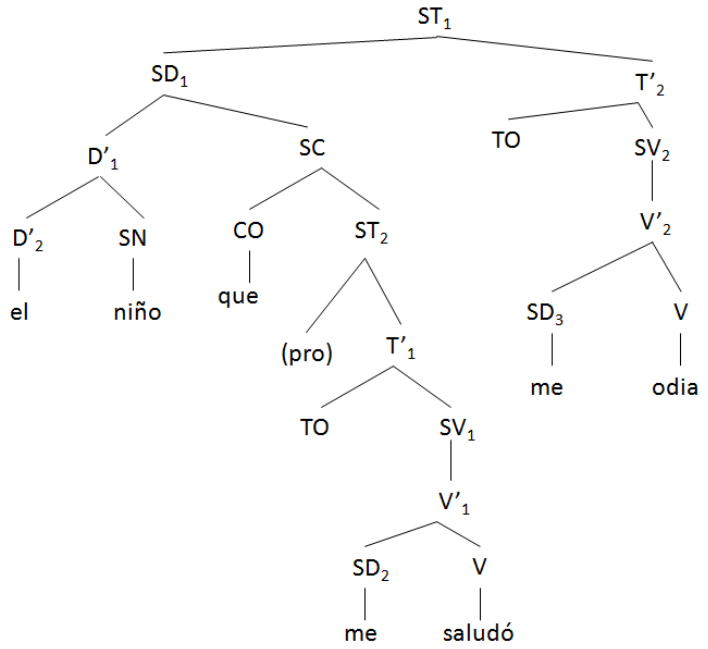


Figura 3.1: Ejemplo de un árbol sintáctico para la oración: *El niño que me saludó me odia*

3.2.1. Analizadores sintácticos

Un analizador sintáctico de lenguaje natural es un programa que trabaja con la estructura gramatical de las oraciones. Los analizadores estadísticos utilizan el conocimiento de la lengua adquirido por análisis realizados a mano para tratar de producir el análisis más probable de las nuevas oraciones. Estos analizadores estadísticos todavía cometen algunos errores, pero suelen trabajar bastante bien. Su desarrollo fue uno de los mayores avances en el procesamiento del lenguaje natural en la década de 1990.

Dentro del análisis sintáctico se distingue entre el análisis de constituyentes y el análisis de dependencias. El análisis de constituyentes se caracteriza por el uso de la relación de inclusión (unos sintagmas incluyen a otros y, en el caso básico, se tienen sintagmas compuestos por unidades léxicas). Dada una oración, este análisis construye un árbol sintáctico que es la representación de las relaciones jerárquicas entre los constituyentes sintácticos. En la Figura 3.1 podemos ver un ejemplo de un árbol sintáctico del análisis de una oración. El análisis de dependencias se caracteriza por el uso de relaciones binarias (de dependencia) entre unidades léxicas. Las palabras de una oración dependen unas de otras, así el objeto directo de un verbo depende directamente de él y un adjetivo depende del nombre. El propósito de este

análisis es construir un árbol de dependencias donde se permita representar cada una de las palabras de la oración y donde los arcos entre las palabras representen las dependencias entre ellas. La Figura 3.2 muestra un ejemplo de árbol de dependencias de una oración. El uso de uno u otro depende de distintos factores, entre los que se encuentran el lenguaje con el que se está trabajando, la finalidad y los resultados del trabajo. A continuación presentamos los principales analizadores que trabajan tanto en inglés como en español.

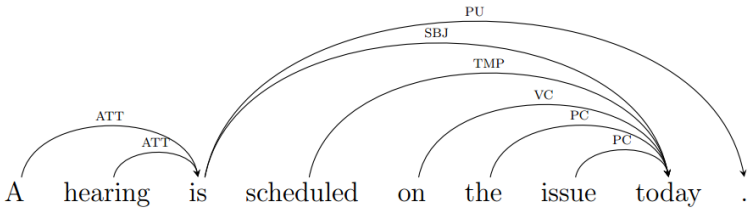


Figura 3.2: Ejemplo de un árbol de dependencias para la oración: *A hearing is scheduled on the issue today.*

3.2.1.1. Analizadores para el inglés

Uno de los primeros analizadores de dependencias para el inglés es Minipar (Lin, 1998). La cobertura del analizador Minipar es bastante amplia. Una evaluación usando el corpus SUSANNE[5] muestra que Minipar alcanza aproximadamente el 88 % de *precision* y el 80 % de *recall* con respecto a las relaciones de dependencia. Es un analizador bastante eficiente que consume pocos recursos. Se puede conseguir una versión ejecutable gratis para uso no comercial.

Entre los analizadores principales para inglés, debido a que realiza tanto análisis de dependencias como de constituyentes, está el analizador desarrollado por la Universidad de Stanford conocido como *Stanford Parser* (Klein y Manning, 2003). Está implementado en Java y la versión original de este programa de análisis fue escrito principalmente por Dan Klein, con código de apoyo de la gramática lingüística desarrollada por Christopher Manning. En versiones posteriores se ha proporcionado una interfaz gráfica de usuario para ver la salida del árbol de estructura de frase del analizador. Además de proporcionar un análisis de inglés, el analizador puede ser y ha sido adaptado para trabajar con idiomas distintos del inglés. Por ejemplo, incluye un analizador de chino basado en el Treebank chino, un analizador de alemán basado en el corpus Negra y un analizador de árabe conforme al Penn Arab

[5] http://www.grsampson.net/SueDoc.html

Treebank. También se ha utilizado para otros idiomas, como italiano, búlgaro y portugués. El analizador proporciona una salida de dependencias, así como la estructura de la frase en árboles de constituyentes.

3.2.1.2. Analizadores para el español

Entre los analizadores para el español destacamos el analizador de dependencias JBeaver (Herrera et al., 2007). Fue desarrollado utilizando Maltparser (Nivre, 2003), un sistema de análisis de dependencias basado en datos que se puede utilizar para inducir un modelo de análisis de datos y para analizar los nuevos datos utilizando un modelo inducido. JBeaver se caracteriza por ser autónomo, fácil de instalar y de utilizar, mediante interfaz gráfica o por comandos de consola, y además tiene una elevada precisión. Con JBeaver se pueden crear corpus de entrenamiento, entrenar a un sistema automático de aprendizaje y realizar análisis y evaluaciones de manera tanto estadística como gráfica.

Como analizador más conocido y utilizado para el español, destacamos el analizador FreeLing desarrollado por la Universidad Politécnica de Cataluña, en el centro de investigación TALP[6]. FreeLing (Padró et al., 2010) es una biblioteca desarrollada para la prestación de servicios de análisis del lenguaje. La versión actual permite identificación del lenguaje, tokenización, división en oraciones, análisis morfológico, reconocimiento de entidades nombradas y clasificación, reconocimiento de fechas, números, magnitudes, ratios, codificación fonética, análisis sintáctico superficial, análisis de dependencias, desambiguación, y resolución de correferencias. Se espera que en versiones futuras se mejore el rendimiento en las funcionalidades existentes, así como la incorporación de nuevas características. FreeLing está diseñado para ser utilizado como una biblioteca externa de cualquier aplicación que requiera este tipo de servicios.

3.2.2. GATE

La herramienta GATE (General Architecture for Text Engineering) (Cunningham et al., 2002) tiene como filosofía reusar, no reinventar, por lo que sus objetivos principales son integrar e interoperar con otros sistemas y herramientas específicas ya existentes.

Tiene una interfaz gráfica y está integrado en un entorno de desarrollo que facilita las diferentes tareas para procesar y editar documentos. GATE es de libre acceso y el software de procesamiento de lenguaje utiliza estructuras de datos y algoritmos especializados tales como gráficos de anotación o máquinas de estados finitos.

La arquitectura de GATE permite que los elementos del sistema de pro-

[6] http://www.talp.upc.edu/

cesamiento de lenguaje natural se puedan dividir en varios tipos de componentes, llamados recursos. Estos recursos son reutilizables en otras interfaces bien definidas. Se definen tres tipos de componentes:

1. *LanguageResources (LRs)*: representan entidades como lexicones, corpus y ontologías.
2. *ProcessingResources (PRs)*: representan entidades que principalmente son algorítmicas, como son parsers y generadores.
3. *VisualResources (VRs)*: representan componentes de visualización y edición que participan en la interfaz gráfica.

El conjunto de recursos integrados en GATE es conocido como CREOLE (*a Collection of REusable Objects for Language Engineering*). Todos los recursos pueden ser exportados como un fichero *Java Archive* (.JAR) más un archivo de configuración en XML. Cuando un conjunto de recursos ha sido desarrollado, éstos se pueden incluir en una aplicación cliente usando *GATE Embedded*. GATE trabaja con varios formatos de documentos incluidos XML, RTF, *email*, HTML, SGML y texto plano. En todos los casos el formato es analizado y convertido en un modelo sencillo unificado de anotación, generando un documento GATE. Los documentos GATE, los corpus y las anotaciones son almacenados en bases de datos y pueden ser visualizados en el entorno de desarrollo.

GATE ayuda a la creación de estas estructuras complejas, a la visualización de los resultados de procesamiento y a la medición de su precisión según los resultados producidos manualmente o semi-automáticamente. En la Figura 3.3 podemos ver la interfaz de GATE, con las distintas aplicaciones, recursos del lenguaje y recursos de procesamiento para trabajar sobre un texto dependiendo de los objetivos que se tengan.

3.3. Herramientas específicas para el tratamiento de expresiones numéricas

En el caso de simplificar expresiones numéricas, se necesitan herramientas específicas que permitan procesar este tipo de información para su posterior tratamiento. En esta sección presentamos las distintas herramientas específicas utilizadas en el trabajo de la tesis. Hemos revisado dos herramientas para el inglés: un parser específico para analizar y anotar las expresiones numéricas presentes en el texto, y un programa específico de aproximación de proporciones que nos permite obtener los posibles candidatos de simplificación dada una expresión de entrada. Para el español, revisamos la herramienta específica JAPE (*Java Annotation Patterns Engine*) perteneciente a GATE que nos ha permitido definir expresiones regulares para anotar las expresiones numéricas de los textos.

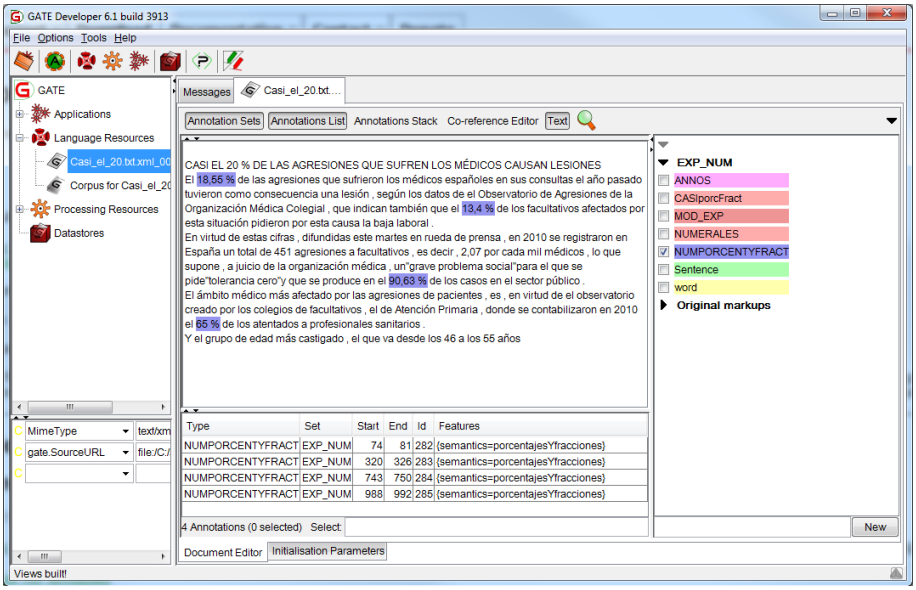


Figura 3.3: Ejemplo de la interfaz de GATE para el procesamiento de un texto

3.3.1. Analizador de expresiones numéricas en inglés

Sandra Williams desarrolló un sistema que combina sintaxis y semántica para analizar y extraer las expresiones numéricas de textos en inglés (Williams, 2010). Así, el sistema funciona como un modelo teórico de cómo las expresiones numéricas son organizadas sintácticamente y un módulo de extracción de información que realiza anotaciones semánticas en formato XML de los modificadores y cantidades que forman las expresiones del texto.

El programa está escrito en Java y reconoce y anota expresiones numéricas en un texto. Recibe como entrada un texto plano y genera como salida un archivo en formato XML con las oraciones anotadas y las expresiones numéricas delimitadas con las etiquetas *<numex...>* y *</numex>*. Utiliza gramáticas con reglas en formato BNF.

El analizador funciona realizando el siguiente proceso:

1. Lee un texto plano que recibe como entrada.
2. Divide el texto en oraciones.
3. Divide cada oración en palabras.
4. Recorre las palabras de una oración con una ventana de seis palabras.
5. Analiza los modificadores numéricos, si existen.
6. Analiza los porcentajes.

7. Analiza las fracciones.
8. Analiza cardinales y números decimales.
9. Analiza cantidades monetarias.
10. Analiza las unidades, si las hay.
11. Genera un archivo de salida con anotaciones en formato XML.

Mostramos un ejemplo de texto de entrada con las expresiones numéricas marcadas en negrita.

> ***Maths and science comeback as A-Level grades soar***
> *A record number of students passed A-levels this year and more achieved A grades than ever before as the Government promised make the qualification tougher. The Joint Council for Qualifications published* ***827,737*** *grades for A-level this year, up from* ***805,657*** *in* ***2007****. A grades went to* ***25.9 per cent*** *of the entries, up from* ***25.3 per cent*** *- and in Northern Ireland more than a third of students achieved an A. Girls continue to outshine boys at grades A-E, but the gap is beginning to narrow - down* ***0.3 per cent*** *at grade A. Entries for maths rose* ***7.5 per cent*** *from 2007, to* ***65,239****, while further maths was up* ***15.5 per cent****, to* ***9,483*** *entries. Less traditional subjects continued to increase in popularity with Chinese, Arabic and Russian showing steady increases every year since 2002. Some other languages suffered with a decrease in the number of students taking German, down* ***0.9 per cent*** *from 2007. But the number sitting French went up by* ***2.8 per cent*** *and there as a* ***1.5 per cent*** *rise in the number opting for Spanish. Sciences also fared well with entries for chemistry up* ***3.5 per cent****, physics up by* ***2.3 per cent*** *and biology up by* ***2.7 per cent****. Among the subjects showing increases were the sciences with entries for chemistry up* ***3.5 %****, biology up* ***2.7 %*** *and physics up* ***2.3 %****. Dr Jim Sinclair, director, JCQ, said the record results were a cause for celebration. "These results are excellent and we congratulate all students on their achievement. The results show not only an improvement in the grades achieved but also an increased entry for mathematics, sciences and languages, which are positive and encouraging developments all round."*

A continuación, podemos ver parte del archivo XML generado como salida, donde están las expresiones numéricas anotadas con las etiquetas que genera el analizador.

```
<doctype w3c-doctype="numgen">
<header>
<title>
    Example XML markup for numerical expressions
</title>
</header>
<article id="1" topic="Maths and science comeback as A-Level
grades soar" date="14-Aug-08">

<sentence id="2">
    A record number of students passed A-levels this year
    and more achieved A grades than ever before as the
    Government promised make the qualification tougher.
</sentence>

<sentence id="3">
    The Joint Council for Qualifications published
    <numex id="1" type="ordinal" digits="yes" value="827,737">
        827,737
    </numex>
    grades for A-level this year, up from
    <numex id="2" type="ordinal" digits="yes" value="805,657">
        805,657
    </numex>
    in
    <numex id="3" type="date" digits="yes" units="year"
    value="2007">
        2007
    </numex>.
</sentence>

<sentence id="4">
    A grades went to
    <numex id="4" type="percentage" digits="yes" value="0.259">
        25.9 per cent
    </numex>
    of the entries, up from
    <numex id="4" type="percentage" digits="yes" value="0.253">
        25.3 per cent
    </numex>
    - and in Northern Ireland more than a third of
    students achieved an A.
</sentence>
...
</doctype>
```

3.3.2. Programa de aproximación de proporciones en inglés

Dada una proporción (un valor de entrada entre 0 y 1), este programa genera un conjunto de versiones equivalentes indicando el tipo matemático de las mismas (fracciones (F) o porcentajes (P)), la relación, el valor representado en fracción y el tipo de modificador que se puede usar junto con su valor en porcentaje o fracción según corresponda.

El desarrollo de este programa fue parte del proyecto *NumGen*. Es un modelo formal para planificar especificaciones para proporciones (números entre 0 y 1) y está formulado a base de restricciones lógicas. Usan gramáticas de generación para expresar en lenguaje natural las distintas soluciones que genera a partir de la proporción de entrada.

En la Figura 3.4 podemos ver un ejemplo de la salida del programa para una proporción de entrada dada. El funcionamiento más detallado y el diseño del programa están descritos en el trabajo de Power y Williams (2012).

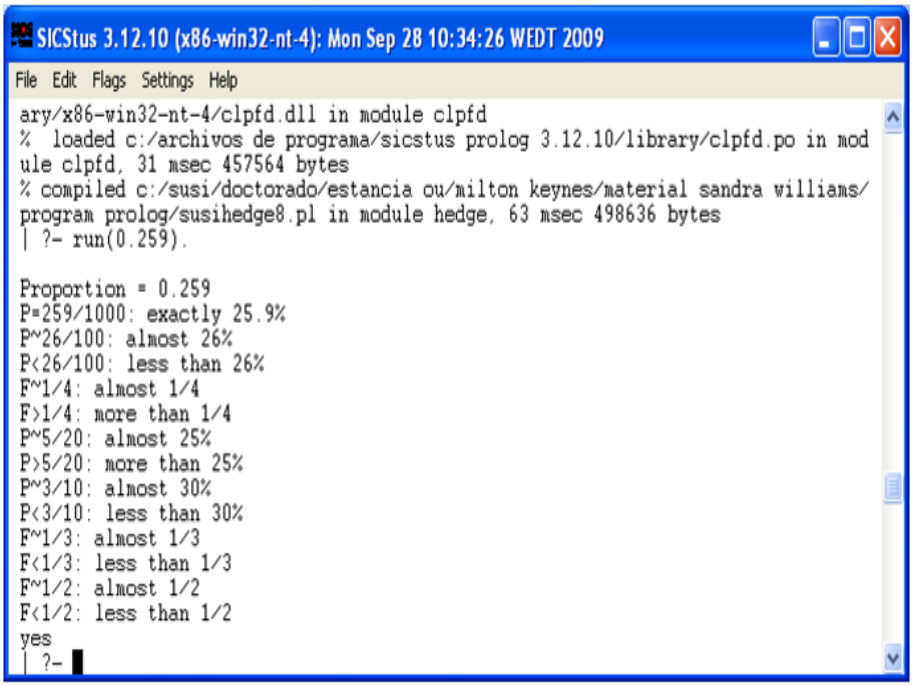


Figura 3.4: Ejemplo de salida del programa de aproximación de proporciones

3.3.3. JAPE (Java Annotation Patterns Engine)

JAPE (*Java Annotation Patterns Engine*)[7] pertenece a GATE y reconoce expresiones regulares implementadas en GATE en documentos anotados.

[7] https://gate.ac.uk/sale/tao/splitch8.html#x12-2170008

JAPE es una versión de CPSL- *Common Pattern Specification Language*[8].

Las gramáticas JAPE consisten en un conjunto de fases, cada una de las cuales tiene un conjunto de patrones y reglas. Estas fases se ejecutan secuencialmente y constituyen una cascada de estados finitos sobre anotaciones. El lado izquierdo de la regla (*Left-hand-side, LHS*) está formado por un patrón de anotación. El lado derecho de la regla (*Right-hand-side, RHS*) consiste en las sentencias de manipulación de la anotación. Las anotaciones del lado izquierdo pueden ser referenciadas en las reglas del lado derecho, usando las etiquetas definidas en los elementos del patrón. La parte izquierda de la regla JAPE es lo correspondiente a lo que precede al símbolo "–>", y la parte derecha a lo que le sigue. Cuando la parte izquierda coincide con la anotación de un documento GATE, entonces el lado derecho especifica lo que se tiene que hacer con el texto correspondiente. Consideremos el siguiente ejemplo de una regla simple, con la que se quiere etiquetar con la etiqueta *"Sport"* una parte del texto que ha sido anotada con un patrón definido anteriormente:

Phrase: category
Input: Lookup
Options:
Rule: SportsCategory
({Lookup.majorType == "Sports"}
):labelS
– – >
:labelS.Sport={rule="SportsCategory"}

La primera línea presenta la *JAPE grammar* bajo la etiqueta *category*. Las anotaciones de entrada tienen que ser también definidas al comienzo de cada gramática. En este caso, la anotación es *Lookup*. También se pueden añadir opciones. Después se define el nombre de la regla tras la etiqueta *Rule*, en nuestro caso *SportsCategory*. A continuación empieza el lado izquierdo de la regla, con el que le estamos diciendo que encuentre una anotación con el patrón *Lookup.mayorType == "Sports"* y temporalmente, le ponga la etiqueta de *labelS*. Al otro lado de la regla, a partir del símbolo "–>", se indica que cuando se encuentre la etiqueta temporal *labelS* sea renombrada por la etiqueta *Sport* y anota como propiedad el nombre de la regla que se ha aplicado (*rule="SportsCategory"*).

Hemos mostrado un ejemplo sencillo de definición de una regla JAPE, pero usando la sintaxis y las reglas de definición de JAPE se pueden definir patrones muy complejos, que consiguen anotar los documentos con las etiquetas que se les indique.

[8]Una buena descripción de la versión original de este lenguaje está en http://www.ai.sri.com/ appelt/TextPro/

Resumen y conclusiones

En este capítulo hemos presentado diferentes herramientas y recursos que se pueden usar en la tarea de simplificación de textos. Hemos presentado el corpus como recursos de simplificación, hemos revisado diferentes herramientas de análisis de texto y herramientas específicas necesarias para el tratamiento de información numérica.

Para el trabajo presentado en esta tesis se han seleccionado corpus específicos para cada lenguaje que nos han permitido trabajar a nivel de expresiones numéricas. Se ha usado el analizador diseñado por Sandra Williams para la parte del trabajo que utiliza textos en inglés. La decisión de usar este analizador para el inglés se basa en que el análisis que nos proporcionaba el Stanford Parser no identificaba las expresiones numéricas con la finalidad que se buscaba. Por ejemplo, los modificadores y las unidades de la expresión no se consideraban bajo el mismo subárbol del análisis.

Para la parte en español del trabajo presentado en esta tesis, hemos utilizado FreeLing para realizar el análisis sintáctico de los textos en español con los que se trabaja, centrándonos en el análisis de las expresiones numéricas.

En la propuesta para el inglés, hemos utilizado el recurso específico para el tratamiento de proporciones. En la propuesta para el español, hemos usado GATE y en concreto las gramáticas JAPE para el tratamiento de la información numérica.

En el siguiente capítulo, presentamos las bases teóricas para la simplificación de textos. Además de un modelo genérico, presentamos la identificación experimental de las estrategias de simplificación de expresiones numéricas realizada para decidir qué tipo de transformaciones teníamos que implementar en nuestra aproximación automática.

Capítulo 4

Bases teóricas para la simplificación de textos centrada en expresiones numéricas

Como ya hemos visto en la introducción de esta tesis, la necesidad de simplificar textos para las personas que tienen dificultades de comprensión por una razón u otra, es una realidad palpable. Las iniciativas desarrolladas para generar manualmente textos simplificados suponen un coste y esfuerzo que no resulta útil por la cantidad de información cambiante que manejamos hoy en día.

Dentro del proceso de simplificación de textos hay un gran abanico de opciones para llevar a cabo distintos tipos de simplificación. Hay diferente información que hay que considerar en cada etapa, dependiendo de los objetivos marcados y de la finalidad con la que se realiza la simplificación textual. En nuestro trabajo nos centramos en la simplificación de expresiones numéricas para ayudar a leer y a comprender un texto con alta carga de información numérica.

Uno de los principales objetivos del trabajo que presentamos en esta tesis es el desarrollo de un modelo computacional para la simplificación automática de expresiones numéricas y las variables que hay que considerar para adaptarlo en cada caso. Para ello hemos estudiado, analizado y decidido qué tipo de operaciones necesitamos implementar y hemos llevado a cabo una identificación experimental de las estrategias de simplificación que utilizan los humanos a la hora de adaptar las expresiones numéricas presentes en un texto.

Como ya definimos en la introducción del trabajo, consideramos una *expresión numérica* como una expresión que representa a una cantidad, como

53 % o *3489*, opcionalmente modificada por un modificador numérico como *más de un cuarto* o *alrededor de 97 %* y opcionalmente acompañada por unidades como *kms*, *litros* o *gramos*. Además, consideramos como estrategias de simplificación a las distintas transformaciones que se utilizan en el proceso de simplificación manual. Por ejemplo, que se cambie la representación matemática de la expresión usando fracciones en lugar de porcentajes, o que se usen modificadores numéricos cuando se redondea la cantidad original.

Nuestras metodologías tiene dos referencias fundamentales. Por una parte, las "*Directrices para materiales en lectura fácil*" de la IFLA, y por otra, las pautas tituladas "*El camino más fácil*", publicadas por la Asociación Europea ISLMH (hoy en día, *Inclusion Europe*), que se basan en las pautas de la IFLA y aportan más matices. Tanto la IFLA como *Inclusion Europe* son conscientes en que las pautas que ellos proponen no deben considerarse de forma dogmática en su aplicación, sino que es preferible la flexibilidad según el tipo de texto, el público objetivo al que se dirige y el idioma del texto.

Somos además conscientes de que en el proceso de simplificación de expresiones numéricas planteamos disminuir la dificultad numérica a costa de aumentar la dificultad sintáctica. Es decir, haciendo una transformación sintáctica, conseguimos una expresión numérica más simple a nivel de comprensión matemática, pero la construcción sintáctica se modifica, y normalmente aumenta en componentes, lo que produce una estructura sintáctica más compleja. Por ejemplo, imaginemos que la expresión matemática original *26,2 %* se simplifica por *más del 25 %*. En el proceso de simplificación se ha disminuido la dificultad numérica, ya que se ha redondeado la cantidad original, perdiendo así la precisión de la original, y se ha aumentado la dificultad sintáctica, ya que se ha añadido un modificador que complica la estructura sintáctica original. Sabemos que la pérdida de precisión no conlleva problemas si no, al contrario, facilita el recordar y comprender mejor los datos numéricos presentes en el texto, y que el cambio en la estructura sintáctica no conlleva una dificultad extrema que complique el acceso a la información (McCloskey et al., 1985).

Las bases teóricas que presentamos en este capítulo son fundamentalmente dos: primero un modelo genérico para la simplificación de textos y un modelo específico para el tratamiento de la información numérica, y segundo, las estrategias de simplificación que queremos automatizar identificadas en distintos estudios experimentales diseñados con diferentes metodologías para dos lenguajes concretos, el inglés y el español.

4.1. Descripción y etapas del modelo genérico para la simplificación de textos

En esta sección presentamos las diferentes etapas de nuestro modelo y cuál es su labor, partiendo de un texto original que se quiere simplificar.

La Figura 4.1 muestra las etapas del modelo. Podemos observar que hay cuatro etapas principales en el proceso de simplificación de un texto, que luego veremos con detalle. Además, hay diferentes variables que determinan la configuración del modelo en cada etapa.

En nuestro modelo consideramos cinco variables principales que entran en juego en distintas etapas del proceso:

1. El lenguaje del texto original, ya que éste afecta a todas las etapas del modelo y determina las herramientas y recursos que pueden ser usadas para analizar el texto, así como las operaciones de simplificación que pueden ser aplicadas.

2. La unidad de descomposición a partir de la cual se va a llevar a cabo la descomposición del texto en la etapa correspondiente. Esta unidad puede ser el párrafo, las oraciones o las palabras, entre otras.

3. El tipo de texto que estamos tratando, como por ejemplo, noticias, recetas, informes u otros.

4. El usuario final para el que se está llevando a cabo la simplificación del texto original, bien sean niños, personas mayores, personas con discapacidad cognitiva o personas que están aprendiendo una lengua.

5. El nivel de dificultad al que se quiere adaptar el texto final, ya que según este nivel las transformaciones que se aplicarán serán unas u otras.

A continuación vamos a ver con más detalle cada una de las etapas de este modelo.

4.1.1. Etapa 1: Análisis del texto

Esta primera etapa de nuestro modelo recibe como entrada el texto plano que se quiere simplificar. La variable que entra en juego en esta etapa es el lenguaje del texto con el que estamos trabajando, ya que determina las herramientas que van a ser usadas para el análisis. La salida de esta etapa es el texto analizado.

Dependiendo de los objetivos específicos que se quieran conseguir en el proceso de simplificación de texto, diferentes tipos de análisis del texto de entrada pueden ser necesarios. En términos generales, la mayoría de los sistemas aplican pasos básicos en el procesamiento del lenguaje natural como son: la tokenización, el separar el texto en oraciones, el etiquetado de las categorías gramaticales (*part-of-speech tagging*) y el análisis sintáctico del texto.

Tokenización: Cada oración tiene que ser separada en los *tokens* que la forman. Este proceso también conlleva algunas dificultades como son

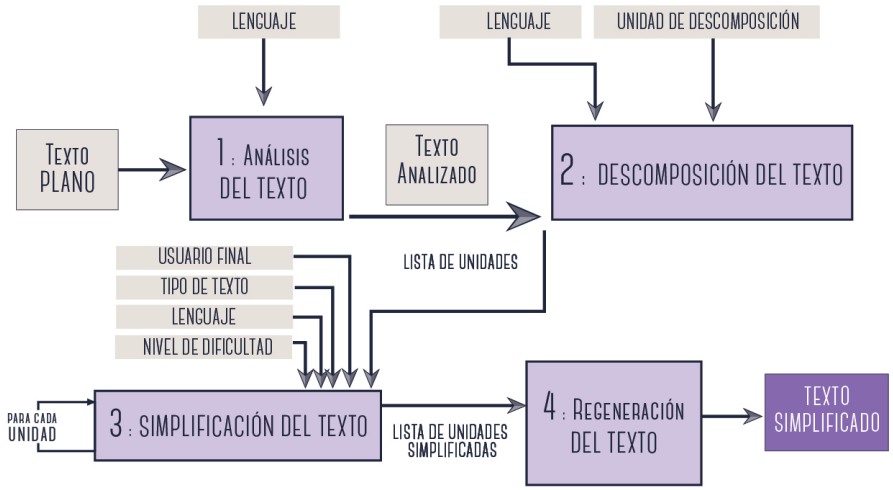


Figura 4.1: Etapas del modelo genérico de simplificación automática de textos

las contracciones en inglés *It's blue* (*It is blue*) o *She's the flu* (*She has the flu*), o en español cuando se combinan varios morfemas en una unidad simple interpretada como una única palabra, como es el caso de la unificación *de* más *el* dando el *token del*. Existen diferentes herramientas y recursos ya implementados para llevar a cabo esta tarea.

Separación en oraciones: Este tipo de análisis a veces se realiza en distintas etapas, pero la idea principal es que el texto de entrada tiene que ser separado en las oraciones que lo constituyen. Hay algunos problemas obvios que nos podemos encontrar a la hora de llevar a cabo la separación automática de las oraciones, como es el caso de la presencia del signo de puntuación (.) cuando no separa oraciones (en abreviaturas o acrónimos), o el caso de las oraciones que forman los titulares de una noticia donde la oración no acaba en punto. Los analizadores sintácticos tienen métodos implementados para realizar la separación en oraciones de un texto dado, utilizando reglas sobre los *tokens* para determinar la separación de las oraciones.

Etiquetado de categorías gramaticales (Part-of-Speech Tagging): A cada palabra se le asigna su categoría gramatical (es decir, verbo, nombre, adjetivo, etc.), y algunas veces también ciertos atributos expresados a través de la flexión de la palabra como son el género, número o tiempo. Diferentes métodos pueden ser usados para llevar a cabo esta tarea, principalmente métodos secuenciales, como son los Modelos Ocultos de Markov, los árboles de dependencias o las gramáticas regu-

lares. Existen diferentes herramientas que implementan estos métodos para realizar esta tarea.

Análisis sintáctico: La sintaxis del lenguaje natural es la forma en la que las palabras individuales son combinadas para formar unidades más complejas. Por un lado, la sintaxis define qué oraciones son gramaticalmente correctas y cuáles no, y por otro lado, influye en la interpretación semántica. El objetivo del análisis es llevar a cabo el etiquetado del texto, asignando a cada palabra su categoría gramatical y definiendo las relaciones que hay entre ellas.

4.1.2. Etapa 2: Descomposición del texto

En esta etapa la entrada es el texto analizado obtenido de la etapa anterior. Para conseguir el objetivo de esta etapa dos variables son consideradas: el lenguaje con el que estamos trabajando y la unidad de descomposición considerada. La finalidad de esta etapa es descomponer el texto en las unidades que van a ser el objetivo del proceso de simplificación, tales como palabras, oraciones o párrafos. Las operaciones de simplificación de la siguiente etapa pueden ser aplicadas a la unidad entera o alguna de sus partes, según los objetivos de simplificación que se planteen.

Aunque se trata de una tarea bastante sencilla, es de suma importancia para el resto del proceso. Esta etapa también puede implicar una tarea de selección de las unidades objetivo. Por ejemplo, si el objetivo del proceso de simplificación consiste en sustituir las palabras difíciles por otras más sencillas, en esta etapa solo las palabras difíciles serán identificadas del texto de entrada como unidades objetivo en el proceso de simplificación.

4.1.3. Etapa 3: Simplificación del texto

La entrada de esta etapa es la lista de unidades de descomposición obtenidas del texto en la etapa anterior. En esta etapa del proceso cuatro variables tienen que ser consideradas: el nivel de dificultad al que se quiere adaptar el texto, el lenguaje con el que estamos trabajando, el tipo de texto y el usuario final para el que se está simplificando. Las cuatro variables determinarán qué operaciones de simplificación son necesarias y deben ser aplicadas en el proceso de simplificación del texto original.

Hay diferentes estrategias posibles de simplificación, como son las transformaciones sintácticas, donde la estructura de una oración, o parte de ella, es transformada, las sustituciones léxicas, donde solo ciertas palabras son modificadas, la eliminación de información no necesaria o la inserción de información que ayude a comprender el texto. De esta manera, una serie de transformaciones se aplican a cada unidad lingüística del texto original para obtener la correspondiente unidad simplificada, que formará parte de la

versión simplificada final.

La salida de esta etapa es una lista de unidades de descomposición simplificadas, obtenidas como resultado de aplicar las diferentes transformaciones correspondientes a las operaciones de simplificación necesarias a cada unidad del texto original.

4.1.4. Etapa 4: Regeneración del texto

En esta etapa final lo único que queda por hacer es recomponer el texto. Puede ser mediante la elaboración de las versiones simplificadas de las unidades que son el resultado de las etapas anteriores, o, si se produjo un proceso de selección durante la descomposición textual, usando las versiones simplificadas de las unidades objetivo en combinación con el resto del texto de entrada. De este modo se construye una versión simplificada del conjunto, obteniendo así un texto simplificado como resultado final del sistema de simplificación.

4.1.5. Combinación de varias estrategias de simplificación

En algunos casos puede ser necesario combinar más de un enfoque de simplificación para lograr el resultado deseado. Cuando varias estrategias de simplificación van a ser aplicadas, hay que definir un arbitraje de actuación para decidir el orden en el que se tienen que aplicar sobre el texto.

Las combinaciones de enfoques radicalmente diferentes - por ejemplo, cuando las técnicas de resumen basadas en la extracción de oraciones completas son combinadas con simplificaciones léxicas o sintácticas dentro de las oraciones - también pueden requerir diferentes instancias de la etapa 1: análisis de texto (sección 4.1.1).

Otro ejemplo es la combinación de la sustitución de las palabras difíciles por otras más fáciles con la reescritura de las construcciones sintácticas complejas por otras más simples. En estos casos, cada enfoque diferente requiere una instanciación de la etapa 2 de descomposición del texto (sección 4.1.2) - para identificar y seleccionar las unidades objetivo para el enfoque de simplificación concreto - y de la etapa 3 de simplificación del texto para aplicar las transformaciones concretas necesarias en cada caso.

Ciertos tipos de estrategias de simplificación pueden involucrar eliminación o inserción de información, como las representaciones gráficas del contenido, o las definiciones del diccionario de palabras difíciles o poco frecuentes. En estos casos, la etapa 2 de descomposición del texto (sección 4.1.2) tendría que identificar los elementos concretos a eliminar o los puntos específicos del texto original donde se va a insertar información adicional.

Además, cuando la simplificación prevista requiere añadir información adicional para una unidad concreta, la etapa 3 tendrá que producir la nueva información requerida para ser colocada en la posición que indica la unidad

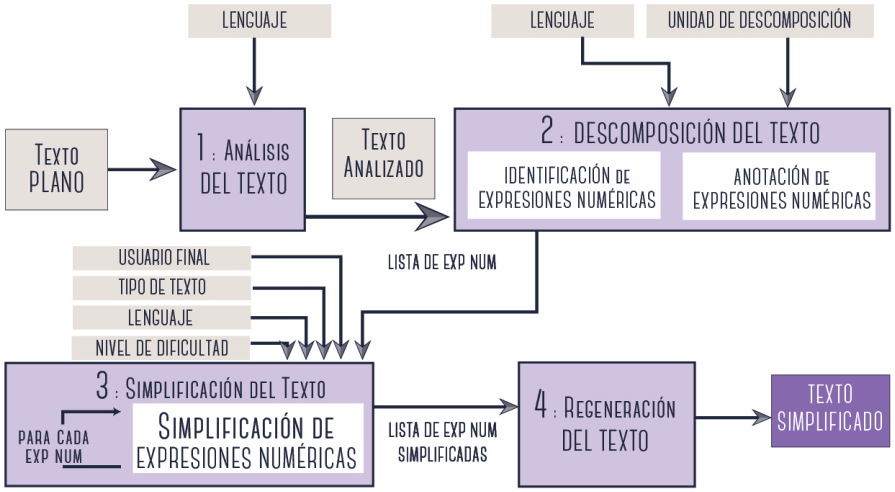


Figura 4.2: Etapas de la instanciación del modelo genérico para la simplificación automática de expresiones numéricas

de descomposición que se está tratando. En estos casos, el contexto y el discurso del texto se tienen en cuenta para que al añadir la información el texto final siga siendo coherente y no se haya perdido, dañado o modificado la información del mismo.

En todos los casos es necesario una compleja instanciación de la etapa 4 de regeneración del texto, para integrar juntos los resultados de los distintos enfoques que se han aplicado al texto original, y obtener así un único texto coherente simplificado como salida final del sistema.

4.2. Instanciación del modelo genérico para la simplificación de expresiones numéricas

En cada caso que se quiera simplificar un tipo de información distinta, hará falta una instanciación del modelo genérico de simplificación de textos que acabamos de presentar en la sección anterior. En nuestro trabajo nos centramos en el tratamiento de la información numérica y para ello el modelo se instancia en un caso concreto para la simplificación de expresiones numéricas presentes en los textos. Además, según los objetivos con los que se trabaje, habrá que decidir qué variables se instancian en el modelo para su futura implementación computacional.

Prestamos atención especial a la etapa 2 (descomposición del texto) del modelo, ya que para llevar a cabo la simplificación de expresiones numéricas descomponemos esta etapa en dos procesos que se corresponden con la iden-

tificación y anotación de las expresiones numéricas de los textos. En la Figura 4.2 podemos ver el modelo de simplificación de expresiones numéricas.

En primer lugar hay que realizar un análisis del texto (etapa 1), utilizando las herramientas que nos sean más propicias para el tratamiento de la información que se va a realizar en las siguientes etapas.

A continuación, en la etapa 2 (descomposición del texto) se realiza la identificación y anotación de las expresiones numéricas presentes en el análisis realizado en la etapa previa. Se identifican diferentes características que pueden formar parte de la expresión numérica para su correspondiente anotación y futura simplificación. La anotación de estas características se realiza aplicando reglas definidas para anotar de manera automática las expresiones.

Dependiendo del lenguaje con el que se trabaje y de las herramientas que se utilicen, la identificación y anotación de las expresiones numéricas se realiza de una manera u otra, siempre basándonos en reglas y gramáticas que nos permitan realizar automáticamente dicho proceso. Para cada uno de los sistemas implementados, veremos las decisiones que se han tomado y como se ha llevado a cabo esta etapa de descomposición del texto.

La etapa 3 (simplificación del texto) se centra en la simplificación de expresiones numéricas. En esta etapa se definen y se implementan las reglas de simplificación de las expresiones numéricas identificadas y anotadas. En la siguiente sección presentamos las metodologías planteadas para la identificación de estas estrategias de simplificación.

Finalmente, en la etapa 4 (regeneración del texto) se obtiene una versión final del mismo con las expresiones numéricas simplificadas. Para ello, se lleva a cabo una sustitución de las expresiones numéricas originales por su correspondiente versión simplificada.

En el capítulo 5 veremos dos sistemas implementados para simplificar expresiones numéricas en inglés y en español, respectivamente, que siguen este modelo de simplificación.

4.3. Metodologías para la identificación de estrategias de simplificación de expresiones numéricas

Con el objetivo de obtener un repertorio de estrategias de simplificación de expresiones numéricas que puedan ser implementadas en un sistema de simplificación automático, se ha realizado una identificación experimental con expertos en el área. Cuando hablamos de estrategias nos referimos a las transformaciones que los expertos utilizan a la hora de realizar el proceso de simplificación manual, para poder generalizar ese tratamiento y automatizarlo para implementar reglas en nuestros sistemas de simplificación.

A continuación, presentamos las dos partes en las que se centra el proceso de simplificación de expresiones numéricas que son las correspondientes a las

dos partes que identificamos en una expresión numérica:

1. Uso de modificadores: una cantidad puede estar o no acompañada por un modificador que determina la precisión de la misma. Por ejemplo, *casi*, *más de* o *alrededor de*.

2. Cantidad: expresión que recoge la información numérica que se transmite. Por ejemplo, *24*, *98 %* o *1/2*.

Opcionalmente una cantidad va acompañada de unidades métricas que pueden variar si cambia también la representación de la cantidad. Por ejemplo, *250 ml* o *1/4 l*. En esta tesis no hemos tratado las unidades en el proceso de simplificación de una expresión numérica.

Proponemos un procedimiento de identificación de las distintas metodologías posibles en una serie de pasos que permitan identificar las estrategias de simplificación de expresiones numéricas. Independientemente del idioma con el que se trabaje, hay que:

1. Plantear unas intuiciones u observaciones de trabajo que queramos validar.

2. Seleccionar los textos con los que vamos a realizar el estudio.

3. Diseñar un estudio donde se presenten distintas opciones de diseño.

4. Analizar los datos recogidos.

Antes de pasar a detallar cada uno de los pasos a seguir en el procedimiento propuesto, expliquemos una serie de conceptos que se utilizan en las intuiciones planteadas:

- Valores comunes y no comunes: los valores de una expresión numérica se clasifican según la frecuencia de uso. Por eso hay valores mucho más conocidos y comunes debido a su alto uso (1/3, 50 %, 1 de cada 4) y valores no tan comunes (1/7, 69 %, 1 de cada 36). Esta frecuencia de uso hace que los valores comunes sean mucho más accesibles para las personas con baja formación numérica.

- Rango central y rangos extremos: normalizando los valores de una cantidad en la escala de 0 a 1, se define el rango central como los valores de 0,2 a 0,8 inclusive y los rangos extremos como los valores de 0 a 0,2 y de 0,8 a 1. Estos rangos permiten clasificar los valores de las expresiones numéricas.

- Modificador: cuantificador que acompaña a la cantidad numérica para expresar su pérdida de precisión.

- Error o pérdida de precisión: diferencia entre el valor exacto de la cantidad y el valor redondeado.

4.3.1. Intuiciones planteadas

Partiendo de lo que hemos aprendido a partir de los trabajos de adaptación de contenidos para personas con dificultades lectoras y de las referencias de las pautas europeas de la IFLA, formulamos las siguientes intuiciones de manera general. Luego en cada caso concreto veremos con detalle las intuiciones planteadas para validarlas a partir de los datos recogidos en cada estudio. Nuestras intuiciones se basan en la elección de la estrategia de simplificación elegida para la versión final, en el uso de modificadores en la versión simplificada y en la pérdida de precisión a la hora de simplificar.

- El valor de la cantidad original (comunes, no comunes, centrales, extremos) influye en la estrategia de simplificación utilizada y en el uso o no de modificador en la versión simplificada.
- La representación matemática de la expresión (dígitos, fracciones, porcentajes, etc.) influye en la accesibilidad de la expresión simplificada.
- La pérdida de precisión influye en la estrategia de simplificación y en el uso de modificadores en la versión simplificada.

4.3.2. Selección del material para el estudio

Para definir el estudio necesitamos un corpus de textos ricos en expresiones numéricas. Esto nos permite seleccionar oraciones con diferentes tipos de expresiones para identificar las distintas transformaciones aplicadas por los humanos para su futura implementación automática. Para cada lenguaje contamos con un corpus específico en el dominio de noticias de prensa, que tienen una alta carga de información numérica. En las secciones correspondientes a los experimentos realizados en distintos idiomas presentamos los detalles del material utilizado en cada caso.

4.3.3. Diseño del estudio

Distintas metodologías pueden ser aplicadas a la hora de diseñar el estudio. Entre las opciones posibles consideramos las siguientes:

- Hacer uso de un corpus paralelo de versiones originales y simplificadas, y analizando dicho corpus identificar las estrategias utilizadas.
- Diseñar una encuesta con expertos para recoger datos y analizarlos para obtener nuevas conclusiones y poder compararlas con nuestras intuiciones de partida.
- Realizar un estudio con usuarios reales que nos permita adaptar el tipo de estrategias de simplificación para ese colectivo concreto.

La finalidad del diseño del estudio es recoger datos que nos permitan identificar un conjunto de estrategias de simplificación de expresiones numéricas utilizadas por los humanos que nos sirva para implementar las transformaciones automáticas en los sistemas de simplificación.

Si en el diseño del estudio se utiliza un corpus paralelo de textos originales y su correspondiente versión simplificada a mano, se identifican las transformaciones aplicadas manualmente a las expresiones numéricas que aparecen en el texto original, consiguiendo así el conjunto de operaciones aplicadas para generar la versión simplificada de los textos del corpus.

El diseño del estudio se puede realizar contando con ayuda de expertos que trabajan a diario con personas con necesidades especiales y están acostumbrados a realizar adaptaciones de textos para estas personas. De ahí que su ayuda sea vital para identificar qué tipo de transformaciones son las más utilizadas por ellos a la hora de simplificar expresiones numéricas del texto original y presentar estas expresiones en la versión simplificada del texto. El diseño se puede realizar usando encuestas donde se le presenta a cada participante un conjunto de oraciones que contienen expresiones numéricas que tienen que simplificar. La idea es seleccionar un rango amplio de valores de las expresiones numéricas para dar mayor cobertura a los distintos tipos de expresiones y obtener así una mayor variedad en las estrategias de simplificación aplicadas.

Otra opción que se puede considerar es diseñar un estudio en el que participe un colectivo de usuarios concretos. Así el diseño del estudio estará adaptado a las necesidades concretas de este grupo de usuarios y se podrán validar nuestras intuiciones de una manera mucho más específica.

4.3.4. Análisis de los datos recogidos

Con los datos recogidos se realiza un análisis para identificar las estrategias utilizadas por los participantes para simplificar las expresiones numéricas y así poder definir las transformaciones que se quieren implementar de manera automática. El análisis se centra en las estrategias utilizadas y en el uso de modificadores.

4.4. Identificación experimental con expertos de las estrategias de simplificación de expresiones numéricas en inglés

Siguiendo el procedimiento propuesto, en esta sección presentamos la identificación experimental de estrategias de simplificación de expresiones numéricas en textos en inglés con la ayuda de expertos.

El abanico de expresiones numéricas es muy amplio, por lo que nos cen-

tramos principalmente en tres tipos de expresiones: fracciones, ratios y porcentajes. Consideramos como estrategias de simplificación el uso de diferentes representaciones matemáticas de la cantidad, acompañadas o no por un modificador.

4.4.1. Intuiciones planteadas para la simplificación de expresiones numéricas en inglés

Nuestras intuiciones han sido formalizadas como hipótesis de trabajo que vamos a validar usando un estudio con expertos.

Planteamos dos hipótesis específicas sobre el uso de estrategias de simplificación, y con respecto al uso de modificadores y la pérdida de precisión en el proceso de simplificación de expresiones numéricas, planteamos cinco hipótesis más. Para nuestras hipótesis consideramos el valor de la expresión numérica normalizado entre 0 y 1, y lo llamaremos proporción.

Hipótesis con respecto al uso de estrategias de simplificación:

H1: Cuando los expertos eligen expresiones numéricas para lectores con baja formación numérica, tienden a preferir redondear a valores comunes de las expresiones, puesto que son más frecuentes. Por ejemplo, medios, tercios y cuartos (*halves, thirds* and *quarters*) normalmente son preferidos frente a otro tipo de fracciones, como octavos o quintos. Y las expresiones como *N in 10* o *N in 100* son preferidas en lugar de otras expresiones como *N in 12* o *N in 96*.

H2: La elección entre las diferentes estrategias de simplificación (fracciones, ratios, porcentajes...) está influenciada por el valor original de la proporción. Los valores pertenecientes al rango central (de *0.2* a *0.8*) y los valores pertenecientes a los rangos extremos (de *0.0-0.2* y *0.8-1.0*) usan diferentes estrategias de simplificación.

Hipótesis con respecto al uso de modificadores y la pérdida de precisión:

H3: El uso de modificadores en la expresión numérica simplificada está influenciado por la estrategia de simplificación considerada.

H4: El uso de modificadores en la expresión numérica simplificada está influenciado por el valor de la expresión numérica original normalizada, con uso de diferentes modificadores para los valores en el rango central (*0.2* a *0.8*) y los valores en los rangos de los extremos (*0.0* a *0.2* y *0.8* a *1.0*).

H5: La pérdida de precisión permitida para la expresión numérica simplificada está influenciada por la estrategia de simplificación seleccionada.

H6: Hay algún tipo de correlación entre la pérdida de precisión y el uso de modificadores, de manera que si crece o decrece la pérdida de precisión influye en la selección del modificador.

H7: Como un caso específico de la hipótesis H6, no se usan modificadores si no hay pérdida de precisión.

4.4.2. Selección del material utilizado para la simplificación de expresiones numéricas en inglés

A partir del material del proyecto *NumGen* (Williams y Power, 2010), se generó un corpus de textos ricos en expresiones numéricas. El corpus está formado por 10 conjuntos de artículos de prensa, 110 textos en total.

Cada conjunto es una colección de artículos del mismo tema donde la información numérica se presenta de forma diferente, lingüística y matemáticamente hablando. Una misma noticia es recogida de distintas fuentes de información y observamos que para una misma información numérica se utilizan distintas representaciones matemáticas.

A continuación mostramos como ejemplo un texto perteneciente al corpus que utilizamos en este estudio. Podemos ver que en un texto formado por 10 oraciones, contando con el título de la noticia, hay un total de 8 expresiones numéricas, en este caso todas ellas porcentajes.

> ***CBI cuts UK growth forecast again***
> *UK economic growth will slow to its lowest level since 1992 next year, employers' group the CBI has warned. In March, the CBI lowered expected GDP growth for 2009 from* ***2.1 %*** *to* ***1.7 %****. It has revised the number downwards once more, now putting expectations at* ***1.3 %****, as households tighten belts due to higher food and fuel prices. The CBI said a "very prolonged period of very sluggish growth" was in prospect for the UK but it was not predicting a recession. The CBI's forecast is well below that of the government, which is still expecting the economy to recover to grow at around* ***2.5 %*** *next year. The Chancellor is expected to address the growing strains on the economy when he delivers his Mansion House speech to the City of London on Wednesday. And just as significantly, the CBI also warned that inflation was likely to breach the governments'* ***2 %*** *for some time to come, driven by higher oil prices. It predicts that inflation will peak at* ***3.8 %****, and says it expects it to rise to* ***3 %*** *when new figures are released on Tuesday. If inflation is* ***1 %*** *higher than the governments' target, then the governor of the Bank of England must write a letter to the Chancellor explaining why he has failed to meet the target.*

4.4.3. Diseño del estudio para la simplificación de expresiones numéricas en inglés

En este caso, nuestro proceso de simplificación sigue una escala de conceptos matemáticos que hemos definido a partir de los niveles de dificultad del curriculum de matemáticas de inglés (*Mathematics Curriculum of the Qualifications and Curriculum Authority*) (Department for Education, 1999). Este documento describe una serie de niveles de enseñanza de las matemáticas y a partir de él, asumimos que los conceptos que se enseñan en los niveles más bajos son más simples que los que se enseñan en los niveles superiores. Siguiendo esta idea hemos definido una escala de conceptos matemáticos para identificar los diferentes niveles de dificultad para comprender conceptos matemáticos. A continuación presentamos la escala definida de menor a mayor dificultad matemática:

1. Expresión numérica representada con palabras (*six*)
2. Expresión numérica representada en números (*600*)
3. Fracciones (*1/4*)
4. Ratios (*1 in 4*)
5. Porcentajes (*25 %*)
6. Porcentajes con decimales (*33.8 %*)

Esta escala es la base para definir los niveles de dificultad considerados en el sistema de simplificación de expresiones numéricas en inglés, que presentamos en el siguiente capítulo (sección 5.1).

Diseñamos nuestro estudio usando una encuesta donde se le presenta a cada participante un conjunto de oraciones que contienen expresiones numéricas que tiene que simplificar. En cada oración aparecen una o más expresiones numéricas marcadas entre corchetes y se le pide que simplifique cada expresión manteniendo el significado de la oración original. En nuestras instrucciones indicamos que las expresiones numéricas pueden ser simplificadas usando cualquier formato: palabras numéricas, dígitos, fracciones, ratios, etc., y que pueden introducir modificadores como *más que*, *casi* o similares si lo creen conveniente. También se indica que el significado de la expresión simplificada debe ser lo más cercano a la expresión original y si lo encuentran necesario, pueden reescribir parte de la oración original o eliminar información redundante.

Nos centramos en simplificar sólo un tipo de expresiones numéricas, los porcentajes, a partir de los cuales hemos identificado las estrategias de simplificación a dos niveles, dependiendo de las expresiones numéricas originales: porcentajes con decimales y porcentajes de números enteros.

Se han elegido tres conjuntos de oraciones candidatas del corpus utilizado para presentárselas a los participantes del estudio: ocho oraciones que contienen sólo porcentajes con decimales y dos conjuntos de ocho oraciones que contienen mezcla de ambos porcentajes, con decimales y sin decimales. Aunque el número de oraciones en cada conjunto es de ocho, el número de expresiones numéricas es mayor, porque hay algunas oraciones que contienen más de una expresión numérica.

Para cada conjunto se presenta un rango amplio de valores de las expresiones numéricas, incluyendo los dos extremos cercanos al *0.0* y el *1.0*. También se incluyen algunas expresiones numéricas que usan modificador numérico. Las oraciones pertenecen a diferentes temáticas del conjunto de textos con el que contamos, intentando dar la mayor cobertura posible con el corpus con el que estamos trabajando.

Para llevar a cabo la identificación experimental de las estrategias de simplificación que usan los expertos a la hora de realizar el proceso manual, planteamos un estudio en el que participaron 34 personas. Los participantes fueron profesores de matemáticas de primaria y secundaria o tutores de adultos con conocimientos básicos de matemáticas, todos hablantes nativos de inglés. La tarea de simplificar expresiones numéricas es difícil, pero es una tarea para la que este grupo está bien cualificado, ya que ellos tienen altos conocimientos de aritmética y están acostumbrados a tratar con personas que no entienden bien los conceptos matemáticos y necesitan adaptación de los contenidos. Conseguimos localizar a este tipo de participantes a través de contactos personales y de mensajes en foros de profesores y tutores de matemáticas en Internet.

El estudio fue presentado a través de la herramienta de encuestas online SurveyMonkey[1], que permite diseñar encuestas de forma sencilla. Cada oración que se les propuso a los participantes tenía marcadas entre corchetes las expresiones numéricas que tenían que simplificar. Detrás de cada oración, cada expresión entre corchetes se mostraba acompañada de una caja de texto donde los participantes escribían su versión simplificada. En la Figura 4.3 se puede ver una parte de nuestro cuestionario.

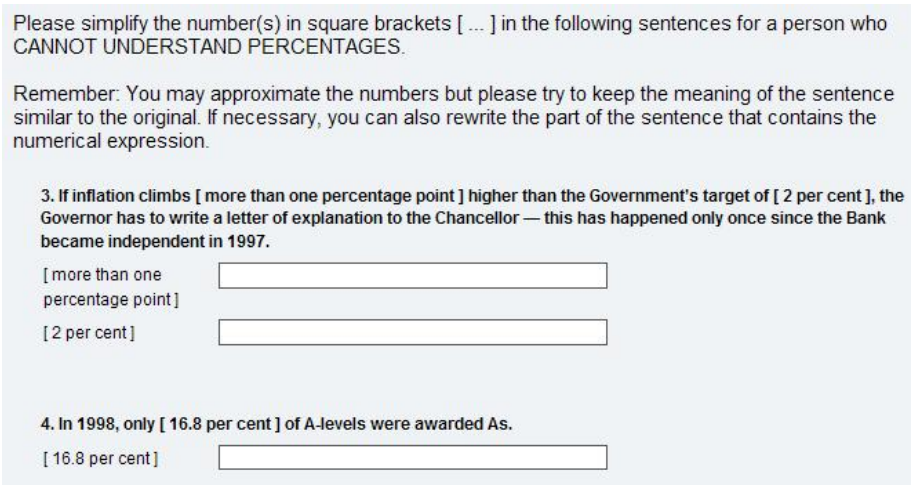
El estudio estuvo dividido en tres partes, como se indica a continuación:

1. Simplificación de expresiones numéricas para una persona que no entiende porcentajes.

2. Simplificación de expresiones numéricas para una persona que no entiende números decimales.

3. Simplificación libre de expresiones numéricas para cualquier persona que tenga problemas de comprensión aritmética.

[1] www.surveymonkey.com


Please simplify the number(s) in square brackets [...] in the following sentences for a person who CANNOT UNDERSTAND PERCENTAGES.

Remember: You may approximate the numbers but please try to keep the meaning of the sentence similar to the original. If necessary, you can also rewrite the part of the sentence that contains the numerical expression.

3. If inflation climbs [more than one percentage point] higher than the Government's target of [2 per cent], the Governor has to write a letter of explanation to the Chancellor — this has happened only once since the Bank became independent in 1997.

[more than one percentage point]

[2 per cent]

4. In 1998, only [16.8 per cent] of A-levels were awarded As.

[16.8 per cent]

Figura 4.3: Parte del cuestionario presentado a los participantes ingleses

Para la parte (2) del estudio, el conjunto de oraciones contiene sólo expresiones numéricas en porcentajes con decimales. Conjuntos mezclados de oraciones con porcentajes en decimales y con porcentajes de números enteros fueron usados para las partes (1) y (3). En todas las partes los participantes podían eliminar parte de la oración o reescribirla si era necesario para su mejor comprensión.

4.4.4. Análisis de los datos para la simplificación de expresiones numéricas en inglés

Dado que teníamos las hipótesis de trabajo divididas según estaban relacionadas con las estrategias de simplificación o con el uso de modificadores, se llevó a cabo el análisis de cada subconjunto a partir de los datos recogidos en la encuesta.

4.4.4.1. Resultados del análisis de las estrategias de simplificación para el inglés

En esta parte del estudio nos centramos en las diferentes estrategias de simplificación usadas por los participantes. Las frecuencias observadas de las diferentes estrategias de simplificación estudiadas (fracciones, ratios, porcentajes y expresiones no numéricas) se muestran en la Tabla 4.1. Los detalles de este estudio se pueden ver en el trabajo de Bautista et al. (2011b).

Primero, con cada bloque de preguntas, un conjunto de estrategias de simplificación fue identificado para cada expresión numérica específica. Estas estrategias fueron agrupadas de acuerdo con la forma matemática y/o expresiones lingüísticas empleadas (fracciones, ratios, porcentajes, no-numérica).

Cuando era necesario, se dividieron de acuerdo con las elecciones de valores numéricos para los constituyentes de la expresión simplificada (denominadores en fracciones, o el valor de referencia en los ratios, por ejemplo).

No todas las estrategias de simplificación ocurren con la suficiente frecuencia para analizarlas con detalle. La solución adoptada en nuestro estudio ha sido agrupar juntas en la subcolumna con la etiqueta *Resto* todas las estrategias de simplificación con una baja frecuencia de uso respecto al total, como reescritura de la oración entera o parte de ella. En el caso de las fracciones, un total de diez diferentes tipos de fracciones fueron usadas por los participantes (cientos, quintos, sextos, etc.), pero solo representamos en las etiquetas de las subcolumnas las más significativas en uso. El resto se agrupó en la subcolumna *OtrasF*. Lo mismo ocurre para el caso de los ratios: los más frecuentes se representan en la tabla, el resto se agrupan bajo la subcolumna *OtrosR*.

Algunos participantes opinaron que algunas oraciones enteras se entenderían mejor si la parte no numérica de la oración también fuera transformada, y en algunos casos que la mejor solución sería eliminar información de la oración directamente.

No Porcentajes (%)

Expresión Numérica	Fracciones				Ratios				No Numérica	Porcentajes	Resto
	Medios	Tercios	Cuartos	OtrasF	Total	N en 10	N en 100	OtrosR	Total				
more than 1 %	3			15	18		6		6	15	18	24
2 %				6	6		12	6	18	3	12	38
16.8 %			3	24	27		15	50	65		9	
27 %		9	71	3	83			12	12		6	
at least 30 %		21	9	12	42	29		6	35		3	9
40 %	21	6		26	53	29			29		6	3
56 %	82				82						6	3
63 %	24	41		9	74	9		15	24		3	
75 %			32		32			29	29	3		24
97.2 %				3	3	3	29	6	38	21	18	12
98 %				6	6		12		12	65	3	9
Media	12 %	7 %	10 %	9 %	38 %	6 %	7 %	11 %	24 %	10 %	7 %	11 %

No Decimales (%)

Expresión Numérica	Fracciones				Ratios				No Numérica	Porcentajes	Resto
	Medios	Tercios	Cuartos	OtrasF	Total	N en 100	N en 1000	OtrosR	Total				
0.6 %	3			3	6	3	6		9	6	47	3
2.8 %				3	3	24			24		47	9
6.1 %						15		3	18		50	3
7.5 %				12	12	3	6	3	12		50	6
15.5 %				15	15	3	6	3	12		44	9
25.9 %			15		15		3	9	12		38	3
29.1 %				3	3	9	3	3	15		50	3
35.4 %		9		3	12	9	3	3	15		41	3
50.8 %	44				44		3		3		21	3
73.9 %			44		44		3	3	6		18	3
87.8 %				3	3	9	3	3	15		47	3
96.9 %				3	3	6	3	3	12		29	12
96.9 %				6	6	9	6	3	18		21	6
97.2 %				3	3	6	6	6	18	3	41	6
97.2 %				3	3	12	3	3	18	3	32	6
98.2 %				3	3	9	3	3	15	6	44	3
Media	3 %	1 %	4 %	4 %	12 %	7 %	3 %	3 %	13 %	1 %	39 %	5 %

Simplificación Libre (%)

Expresión Numérica	Fracciones				Ratios				No Numérica	Porcentajes	Resto
	Medios	Tercios	Cuartos	OtrasF	Total	N en 10	N en 100	OtrosR	Total				
0.7 %							6		6	18	9	26
12 %				6	6	12	3	6	21		21	3
26 %			41		41			12	12			3
36 %		41			41	3		6	9			3
53 %	41				41						6	6
65 %	6	15			21	3	9	6	18		3	12
75 %			15		15			9	9	6	3	15
91 %						21	9		30	6	6	12
above 97 %						3	29		32	12	6	
Media	5 %	6 %	6 %	1 %	18 %	5 %	6 %	4 %	15 %	5 %	6 %	9 %

Tabla 4.1: Frecuencias para las estrategias de simplificación para las tres partes del estudio en inglés: (1) destinado para personas que no entienden porcentajes (No Porcentajes), (2) destinado para personas que no entienden expresiones con decimales (No Decimales) y (3) destinado para personas con baja formación numérica (Simplificación Libre)

Para analizar los resultados que hemos obtenido de nuestro estudio, se llevó a cabo un análisis de la varianza simple (ANOVA) cuyos resultados pueden verse en la Tabla 4.2. Cuando consideramos el estudio completo, última columna, no hay diferencia significativa entre el uso de fracciones, ratios y porcentajes. Sólo el uso de expresiones no numéricas tiene una diferencia significativa con respecto al resto, pero esto es debido a su bajo uso. Sin embargo, cuando analizamos el estudio por partes encontramos resultados interesantes.

Estrategia	**No Porcentajes**			**No Decimales**			**Simplificación Libre**		**Estudio Completo**	
Fracciones	A			A			A		A	
Ratios		B		A			A		A	
Porcentajes			C		B			B	A	
No Numérica			C			C		B		B

Tabla 4.2: Resultados del test ANOVA. Las estrategias que no comparten letra son significativamente diferentes

Como se puede ver en la Tabla 4.1, las fracciones son la simplificación preferida (38 % de uso) para personas que no entienden porcentajes (No Porcentajes). Aunque los participantes utilizaron diez tipos diferentes de fracciones, las más comúnmente usadas fueron medios, tercios y cuartos. La segunda estrategia de simplificación preferida son los ratios (24 % de uso). De los nueve tipos diferentes de ratios empleados (rangos de *N in 10* a *N in 36*), los más comunes fueron *N in 10* y *N in 100*. La siguiente estrategia más usada son las expresiones no numéricas (10 % de uso) para las expresiones originales pertenecientes a los rangos extremos.

Queremos destacar que el 7 % de las expresiones elegidas fueran porcentajes, incluso cuando a los participantes se les preguntó que simplificaran expresiones numéricas para personas que no entendían porcentajes. No estamos seguros si ignoraron nuestras instrucciones, no estaban de acuerdo con ellas o no encontraron otra forma de simplificar la expresión. En cualquier caso, el uso de porcentajes no es significativo con respecto al uso de expresiones no numéricas en esta parte del estudio.

Los porcentajes con números enteros son la estrategia de simplificación preferida (39 % de uso) para las personas que no entienden decimales (No Decimales). Esto refuerza la idea de que son más fáciles de entender que el número original, mientras que al mismo tiempo son la forma más cercana al valor original y la forma matemática más intuitiva de usar. Estamos hablando de casos como cuando la expresión numérica original es *52.4 %* y la expresión simplificada es *a little more than 50 %*, donde se hace uso de un modificador, ya que se pierde la precisión en la cantidad al redondear el número original. Las frecuencias de uso de fracciones (12 %) y ratios (13 %) en esta parte del estudio son muy similares y no son significativamente diferentes. Las

simplificaciones usando expresiones no numéricas fueron muy poco usadas (1 %), a diferencia de en la primera parte del estudio; de hecho, sólo son usadas para puntos periféricos de la escala de la proporción, como son los valores cercanos al 0 o al 1, siendo expresiones del tipo *around none* o *almost all*.

Cuando se pide simplificar libremente (Simplificación Libre) no hay una estrategia de simplificación claramente más usada con respecto al resto de estrategias. El uso de fracciones (18 %) y de ratios (15 %) es similar y ocurre lo mismo para los casos de uso de expresiones no numéricas (5 %) y porcentajes (6 %). Hay un alto uso (9 %) de otras estrategias de simplificación agrupadas en *Resto* en comparación con el resto de partes de la encuesta.

4.4.4.2. Resultados del análisis del uso de modificadores para el inglés

Con miras a utilizar estos datos para diseñar un sistema de simplificación automática, analizamos los resultados de nuestro estudio centrándonos en el uso de los modificadores. Primero, con cada bloque de preguntas, un conjunto de estrategias de simplificación fue identificado para cada expresión numérica específica. Estas estrategias fueron agrupadas de acuerdo con la forma matemática y/o expresiones lingüísticas empleadas (fracciones, ratios, porcentajes). No consideramos las expresiones no numéricas ya que se va a calcular la pérdida de precisión entre las expresiones y para el caso de las no numéricas no se puede calcular. Los detalles de este estudio se pueden ver en el trabajo de Bautista et al. (2011a).

Estos datos pueden ser analizados en términos de pares de una expresión numérica de entrada dada y la expresión simplificada que resulta de la aplicación de una estrategia de simplificación específica. Consideramos tres características importantes en cada pareja identificada:

- La frecuencia de uso de cada una de las representaciones matemáticas disponibles (fracciones, ratios y porcentajes).
- La pérdida de precisión o error involucrada en la simplificación.
- El posible uso de modificador para cubrir esa pérdida de precisión explícitamente en la expresión simplificada.

Para calcular la pérdida de precisión o error, definimos la ecuación 4.1.

$$error = \frac{(ExprNumSimplificada - ExprNumOriginal)}{ExprNumOriginal} \tag{4.1}$$

Se presentan a continuación tres tablas de análisis de cada parte del estudio. En cada tabla se analiza la frecuencia de uso, el error y el uso

Exp. Num.		Frecuencia (%)	Error (%)	Modificador (%)
more than 1 %	Fracciones	18	0	67
	Ratios	6	0	100
	Porcentajes	18	17	50
2 %	Fracciones	6	0	50
	Ratios	18	-1	17
	Porcentajes	12	0	0
16.8 %	Fracciones	26	1	67
	Ratios	65	5	45
	Porcentajes	9	-3	0
27 %	Fracciones	82	-4	86
	Ratios	12	8	75
	Porcentajes	6	6	50
at least 30 %	Fracciones	41	0	93
	Ratios	35	13	67
	Porcentajes	3	0	100
40 %	Fracciones	53	12	50
	Ratios	29	0	10
	Porcentajes	6	0	0
56 %	Fracciones	82	-13	82
	Ratios			
	Porcentajes	6	-5	50
63 %	Fracciones	74	-3	84
	Ratios	24	0	75
	Porcentajes	3	0	0
75 %	Fracciones	32	0	0
	Ratios	29	0	0
	Porcentajes			
97.2 %	Fracciones	3	0	0
	Ratios	38	-8	23
	Porcentajes	18	1	50
98 %	Fracciones	6	0	0
	Ratios	12	0	0
	Porcentajes	3	0	0
Media	Fracciones	39	-1	53
	Ratios	24	2	41
	Porcentajes	7	1	30

Tabla 4.3: Análisis de la frecuencia, la pérdida de precisión y el uso de modificadores en los datos recogidos en la primera parte del estudio de inglés (simplificación para personas que no entienden porcentajes (No Porcentajes)). Todos los valores representados en la tabla son porcentajes

de modificador para las estrategias de simplificación (fracciones, ratios y porcentajes) para cada una de las expresiones numéricas presentadas en cada parte del estudio. La Tabla 4.3 corresponde a la primera parte del estudio (simplificación para personas que no entienden porcentajes). La Tabla 4.4 corresponde a la segunda parte del estudio (simplificación para personas que no entienden decimales) y la Tabla 4.5 corresponde a la tercera parte del estudio (simplificación libre para personas con baja formación).

Para cada expresión numérica de entrada, el conjunto de estrategias de simplificación disponibles se representa como tres líneas en las tablas. Para cada pareja, tres columnas se muestran en la tabla. Las celdas vacías representan que no se utilizó la estrategia. La primera columna presenta la frecuencia relativa de uso con respecto al conjunto total de las posibles estrategias de simplificación utilizadas para esa expresión. La segunda columna captura la pérdida de precisión implicada, representándola en términos del ratio entre el valor numérico original en la expresión de entrada y el valor numérico que se expresa en la expresión simplificada correspondiente (utilizando la ecuación 4.1). Esta relación también se expresa como un porcentaje. La tercera columna indica el porcentaje de expresiones numéricas simplifi-

Exp. Num.		Frecuencia (%)	Error (%)	Modificador (%)
0.6 %	Fracciones	6	25	50
	Ratios	9	22	33
	Porcentajes	47	21	100
2.8 %	Fracciones	3	-29	0
	Ratios	24	6	63
	Porcentajes	47	7	63
6.1 %	Fracciones			
	Ratios	18	-4	50
	Porcentajes	50	-3	82
7.5 %	Fracciones	12	9	75
	Ratios	12	-10	0
	Porcentajes	50	7	41
15.5 %	Fracciones	15	-1	80
	Ratios	12	6	50
	Porcentajes	44	2	33
25.9 %	Fracciones	15	-3	100
	Ratios	12	-3	75
	Porcentajes	38	5	62
29.1 %	Fracciones	3	0	0
	Ratios	15	3	60
	Porcentajes	50	2	71
35.4 %	Fracciones	12	-5	100
	Ratios	15	-4	60
	Porcentajes	41	-1	71
50.8 %	Fracciones	44	-2	93
	Ratios	3	0	0
	Porcentajes	21	0	43
73.9 %	Fracciones	44	1	93
	Ratios	6	1	50
	Porcentajes	18	0	50
87.8 %	Fracciones	3	0	0
	Ratios	15	-1	60
	Porcentajes	47	1	88
96.9 %	Fracciones	3	0	0
	Ratios	12	-2	75
	Porcentajes	29	0	80
96.9 %	Fracciones	6	0	50
	Ratios	18	-1	67
	Porcentajes	21	0	86
97.2 %	Fracciones	3	0	0
	Ratios	18	-1	67
	Porcentajes	41	0	93
97.2 %	Fracciones	3	0	0
	Ratios	18	-1	83
	Porcentajes	32	0	91
98.2 %	Fracciones	3	0	0
	Ratios	15	-2	40
	Porcentajes	44	0	67
Media	Fracciones	11	0	43
	Ratios	14	1	52
	Porcentajes	39	2	70

Tabla 4.4: Análisis de la frecuencia, la pérdida de precisión y el uso de modificadores en los datos recogidos para la segunda parte del estudio de inglés (simplificación para personas que no entienden decimales (No Decimales)). Todos los valores están representados en porcentajes

cadas que contenían un modificador. Todos ellos son valores medios.

Otro punto a explicar es que las frecuencias que pertenecen a la misma expresión no siempre suman el 100 %. Esto se debe a un pequeño número de otros tipos de estrategias de simplificación, como eliminaciones o reescrituras de toda la frase, que no se muestran en la tabla.

En las tres partes del estudio, el porcentaje de simplificaciones que utilizan modificadores es ligeramente mayor que la de aquellos que no utilizan modificadores, especialmente en la segunda y tercera parte del estudio. La adaptación de expresiones numéricas originales añadiendo modificadores representa más del 50 % de los casos. Esto refuerza nuestra hipótesis de que las

Exp. Num.		Frecuencia (%)	Error (%)	Modificador (%)
0.7 %	Fracciones			
	Ratios	6	43	100
	Porcentajes	9	43	100
12 %	Fracciones	6	-17	100
	Ratios	21	-8	71
	Porcentajes	21	-17	100
26 %	Fracciones	41	-4	57
	Ratios	12	-4	50
	Porcentajes			
36 %	Fracciones	41	-8	86
	Ratios	9	-2	67
	Porcentajes			
53 %	Fracciones	41	-6	50
	Ratios			
	Porcentajes	6	-6	50
65 %	Fracciones	21	-5	100
	Ratios	18	-1	33
	Porcentajes	3	0	0
75 %	Fracciones	15	0	20
	Ratios	9	0	33
	Porcentajes	3	0	0
91 %	Fracciones			
	Ratios	29	-1	50
	Porcentajes	6	-1	50
above 97 %	Fracciones			
	Ratios	32	0	64
	Porcentajes	6	2	100
Media	Fracciones	18	-7	69
	Ratios	15	3	59
	Porcentajes	6	3	57

Tabla 4.5: Análisis de la frecuencia, la pérdida de precisión y el uso de modificadores en los datos recogidos para la tercera parte del estudio de inglés (simplificación libre para personas con baja formación (SIMPLIFICACIÓN LIBRE)). Todos los valores están representados en porcentajes

simplificaciones que implican pérdida de precisión se pueden entender mejor si se utiliza un modificador adecuado. Más adelante usaremos los datos de estas tablas para validar nuestras hipótesis de trabajo.

4.4.4.3. Estudio de los modificadores utilizados

Con respecto a los modificadores utilizados, hemos identificado dos posibles papeles que juegan los modificadores como ingredientes de una expresión numérica. En algunos casos los modificadores son usados para indicar que el valor numérico real dado es una aproximación al valor previsto. Usos como *about* o *around* son ejemplos de este caso. Este tipo de modificador se emplea para indicar explícitamente que una cierta pérdida de precisión se ha producido durante la simplificación. En otros casos los modificadores son usados para indicar la dirección en la que el valor simplificado diverge del valor original. Ejemplos de ellos son *under* u *over*. En algunos casos, más de un modificador puede ser añadido en la expresión para indicar tanto la aproximación como la dirección, o para especificar la precisión de alguna manera más concreta en la simplificación, como por ejemplo en el caso de *just under* o *a little less than*.

En nuestro análisis hemos estudiado qué modificadores son los más frecuentes en cada parte de la encuesta. Sólo los modificadores con más de diez

apariciones en total (incluyendo las estrategias de simplificación menos frecuentes y no presentadas en las tablas) han sido considerados en la Tabla 4.6. Hemos observado que las tres partes de la encuesta tienen tres modificadores en común: *about, just over* y *over*. Se utilizan en diferentes estrategias para cada tipo de simplificación. En la segunda parte de la encuesta, donde las simplificaciones de expresiones numéricas eran hechas para una persona que no entendía decimales, es donde más modificadores se han usado, en especial para la estrategia de usar porcentajes. En la última parte de la encuesta, donde hay más libertad para decidir cómo simplificar la expresión numérica original, los participantes usaron menos modificadores comparado con las otras partes.

No Porcentajes			
Modificador	Fracciones	Ratios	Porcentajes
about	15	9	0
at least	8	5	1
just over	21	1	0
more than	9	3	0
over	6	3	2
Total	59	21	3
No Decimales			
Modificadores	Fracciones	Ratios	Porcentajes
about	8	12	6
almost	4	1	8
just over	13	3	39
just under	3	2	27
nearly	7	5	24
over	7	5	9
Total	42	28	113
Simplificación Libre			
Modificadores	Fracciones	Ratios	Porcentajes
about	6	5	1
just over	6	0	5
more than	4	5	0
nearly	4	0	2
over	11	2	3
Total	31	12	11

Tabla 4.6: Uso de los modificadores más frecuentes en cada una de las partes del estudio en inglés.

4.4.4.4. Validación de nuestras hipótesis de trabajo

En esta sección vamos a presentar la validación de nuestras hipótesis de trabajo (sección 4.4.1) a partir de los estudios estadísticos realizados para comprobar si aceptamos o rechazamos cada hipótesis.

Para comprobar la hipótesis H1 (valores comunes o redondeados son preferidos para simplificar el valor de la expresión numérica original) hemos llevado a cabo un estudio estadístico usando una *t-Student* pareada para fracciones y ratios, comunes y no comunes. Los resultados muestran que hay

Estrategia	No Porcentajes		No Decimales	Simplificación Libre	Estudio Completo	
Fracciones	A		A	A	A	
Porcentajes	A	B	A	A	A	
Ratios		B	A	A		B

Tabla 4.7: Resultados del estudio *t-test* ajustado por *la corrección de Bonferroni* para la hipótesis H3 (el uso de modificadores en la expresión numérica simplificada está influenciado por la estrategia de simplificación seleccionada). Las estrategias que no comparten letra son significativamente diferentes

una diferencia estadística significativa entre el uso de fracciones comunes y no comunes analizando las tres partes del estudio por separado y si consideramos el estudio completo (no porcentajes: $p<0{,}001$; no decimales: $p=0{,}07$; simplificación libre: $p<0{,}0001$; estudio completo: $p<0{,}0001$). Sin embargo, en el caso de los ratios, no hay diferencia significativa excepto en el caso de la tercera parte del estudio, correspondiente a la simplificación libre (no porcentajes: $p=0{,}48$; no decimales: $p=0{,}36$; simplificación libre: $p=0{,}006$; estudio completo: $p=0{,}14$). Por lo que nuestra hipótesis es cierta para fracciones, pero no se cumple en el caso de los ratios.

Como se puede ver en los datos recogidos, el uso de tipos diferentes de fracciones parece que depende del valor de la expresión numérica original que se va a simplificar. Los cuartos, tercios y mitades (fracciones comunes) son preferidos en el rango central de *0.2* a *0.8*, y la otra gran variedad de fracciones son usadas en los valores extremos de los valores originales, de *0.0* a *0.2* y de *0.8* a *1.0*. Lo mismo ocurre en el uso de ratios, porcentajes y expresiones no numéricas en los rangos centrales y extremos. Se puede ver por tanto una clara influencia del valor original de la proporción a la hora de elegir entre las diferentes estrategias de simplificación. Esta fue nuestra hipótesis H2 y para validarla, se llevó a cabo un estudio estadístico usando para una *t-Student* pareada para los valores centrales y periféricos de las expresiones numéricas originales para fracciones, ratios, porcentajes y expresiones no numéricas. Los resultados de nuestro estudio estadístico muestran que en el uso de las cuatro estrategias de simplificación usadas (fracciones, ratios, porcentajes y expresiones no numéricas) existe diferencia estadísticamente significativa para valores centrales y extremos de las proporciones de entrada (fracciones: $p<0{,}0001$; ratios: $p=0{,}03$; porcentajes: $p<0{,}0001$; no-numérica: $p<0{,}0001$). El único caso donde no hay diferencia estadísticamente significativa fue en el uso de ratios en la primera parte del estudio (simplificación para personas que no entienden porcentajes) con un *p-valor* de 0,14. Nuestra hipótesis es por tanto aceptada en todos los casos, menos en la primera parte del estudio para el caso de los ratios.

Para comprobar nuestra hipótesis H3 (el uso de modificadores en la expresión numérica simplificada está influenciado por la estrategia de simplifi-

Estrategia	No Porcentajes		No Decimales	Simplificación Libre		Estudio Completo	
Fracciones	A		A	A		A	
Porcentajes	A		A	A	B		B
Ratios		B	A		B		B

Tabla 4.8: Resultados del estudio t-test ajustado por la corrección de Bonferroni para la hipótesis H5 (la pérdida de precisión permitida para la expresión numérica simplificada está influenciada por la estrategia de simplificación seleccionada). Las estrategias que no comparten letra son significativamente diferentes

cación seleccionada), se llevaron a cabo una serie de dos muestras de *t-tests* donde la significancia estadística se ajustó para comparaciones múltiples utilizando *la corrección de Bonferroni*. Los resultados son presentados en la Tabla 4.7. Cuando consideramos los resultados de la encuesta al completo, recogidos en la columna *Estudio Completo*, no hay diferencia significativa en el uso de modificadores en fracciones y porcentajes. Cuando analizamos la encuesta por partes, encontramos resultados similares. No hay diferencia significativa en el uso de modificadores en cualquier estrategia usada en la segunda (no decimales) y tercera (simplificación libre) partes del estudio, pero en la primera parte (no porcentajes) encontramos diferencia significativa entre fracciones y ratios ($p<0.0006$). Estos resultados no apoyan nuestra hipótesis, ya que no demuestran que haya una relación directa entre el uso de modificadores y la estrategia seleccionada.

Se realizó otro *t-test* usando *la corrección de Bonferroni* para analizar el uso de las estrategias de simplificación para los valores centrales y periféricos según la hipótesis H4 (el uso de diferentes modificadores para simplificar expresiones numéricas está influenciado por el valor de la proporción original, con valores en el rango central (*0.2-0.8*) y valores en los extremos (*0.0-0.2* y *0.8-1.0*)). En este caso no hay diferencia significativa. Los resultados muestran que el uso de modificadores no está influenciado por los valores centrales y periféricos, rechazando nuestra hipótesis H4 con un p-valor de $p=0.77$ en el peor caso de la estrategia de uso de porcentajes.

Un nuevo *t-test* usando *la corrección de Bonferroni* se realizó para testear la hipótesis H5 (la pérdida de precisión permitida para la expresión numérica simplificada está influenciada por la estrategia de simplificación seleccionada). La Tabla 4.8 muestra las diferencias significativas entre cada estrategia de simplificación y cada tipo de simplificación. En la columna *Estudio Completo* podemos observar que la pérdida de precisión en fracciones es significativamente diferente a la de ratios y porcentajes. En la primera parte (no porcentajes) hay diferencia significativa entre ratios y el resto de estrategias de simplificación. En la segunda parte (no decimales) no hay diferencia significativa entre ninguna estrategia. Y en la última parte (simplificación

libre) hay solo diferencia significativa entre fracciones y ratios. Estos resultados parecen no apoyar la hipótesis, ya que no hay una relación directa entre el uso de modificadores y la pérdida de precisión en las expresiones numéricas simplificadas.

Para la hipótesis H6 (hay algún tipo de correlación entre la pérdida de precisión y el uso de modificadores) buscamos las correlaciones para cada parte del estudio y para cada tipo de estrategia de simplificación. Se llevó a cabo una medida no paramétrica de la dependencia estadística entre dos variables (la pérdida de precisión y el uso de modificadores) calculada por el *coeficiente de correlación de Spearman*. En general, los resultados muestran que no hay correlación, es decir, que no hay una dependencia lineal entre la pérdida de precisión en la estrategia y el uso de modificadores, rechazando nuestra hipótesis. Por ejemplo, hay casos con una débil correlación (por ejemplo, en la segunda parte de la encuesta para las fracciones tenemos $r=0.49$, $N=17$ and $p=0.03$), y casos donde hay una fuerte correlación (por ejemplo, en la tercera parte de la encuesta, tenemos $r=1$, $N=18$ and $p<0.0001$).

Finalmente, cuando analizamos nuestra hipótesis H7 (si no hay pérdida de precisión no se usan modificadores), hemos trabajado en cada parte de la encuesta para estudiar los casos donde la pérdida de precisión es cero para ver cuál es la tendencia de uso de modificadores.

- En la primera parte de la encuesta (simplificación de expresiones numéricas para una persona que no entiende porcentajes), en un 46 % de las respuestas la pérdida de precisión es cero, y para estos casos solo un 11 % usa modificadores.
- En la segunda parte (simplificación de expresiones numéricas para una persona que no entiende decimales), en un 16 % de las respuestas la pérdida de precisión es cero, y para estos casos solo un 7 % usa modificadores.
- Finalmente, en la última parte (simplificación de expresiones numéricas para una persona con baja formación aritmética en general), en un 23 % de las respuestas la pérdida de precisión es cero, y para estos casos solo un 6 % usa modificadores.

Con estos datos, parece que podemos aceptar nuestra hipótesis H7, es decir, que encontramos evidencia para nuestra suposición de que cuando se simplifican expresiones numéricas, se tiende a no usar modificadores cuando la pérdida de precisión es cero comparando la versión original con la versión simplificada.

Finalmente, en nuestra encuesta había pocos casos donde la expresión numérica original tenía un modificador. Hemos observado que si la expresión numérica original tiene modificador casi siempre la versión simplificada lo mantiene. Hay casos especiales, como por ejemplo *above 97 %* donde no

consideramos el uso de modificador en la versión simplificada, porque en este caso los participantes optaron mayoritariamente por una expresión no numérica que reescriben según el significado como por ejemplo *around all*. En el resto de los casos, el mismo modificador que en la expresión original es casi siempre elegido para la expresión numérica simplificada.

4.4.4.5. Discusión del análisis de los datos para el inglés

Cuando se quiere simplificar para personas que no entienden porcentajes, o para personas con bajo nivel de conocimiento matemático, los participantes de nuestro estudio prefieren usar fracciones para simplificar las expresiones numéricas dadas en el cuestionario, seguidas de ratios como segunda estrategia de simplificación preferida. Cuando se les pide simplificar expresiones numéricas para personas que no entienden decimales, prefieren usar porcentajes con números enteros, redondeando las expresiones numéricas con decimales. Las respuestas recogidas muestran que las fracciones son consideradas como la forma matemática más simple, seguidas de los ratios, pero esto no significa que las fracciones se prefieran a los ratios en todos los casos: el valor de la proporción original también influye en la elección, las fracciones son altamente preferidas para valores centrales (en el rango de *0.2* a *0.8*), y los ratios y las expresiones no-numéricas son preferidas para valores periféricos (anteriores al *0.2* o por encima de *0.8*), siempre dependiendo del tipo de simplificación que se esté llevando a cabo.

Un peligro que se corre en la simplificación de varias cantidades relacionadas es que se puede oscurecer la relación entre ellas. Un caso evidente es redondear dos valores diferentes por un único valor simplificado, es decir, simplificar *48 %* y *52 %*, ambos por *a half*. En algunos de estos ejemplos donde más de una expresión numérica está siendo comparada, algunos de los participantes en nuestro estudio tienden a parafrasear ambas acorde a una base comparable. Esto nos permite pararnos a pensar en el papel que juega el contexto para establecer qué simplificación debe ser usada (el conjunto de expresiones numéricas pertenecientes a una oración dada y el significado del texto donde están). Por ejemplo, las expresiones numéricas correspondientes a los primeros valores de la Tabla 4.1 pertenecen al texto *"If inflation climbs [more than one percentage point] higher than the Governments target of [2 per cent], the Governor has to write a letter of explanation to the Chancellor"*. Como el texto está relacionado con temas monetarios, muchas de las simplificaciones para la expresión *more than one percentage point* fueron expresadas en ratios de la forma *pence in the pound* o *£1 in every £100*.

A través de un estudio llevado a cabo por expertos en conocimientos matemáticos, hemos recogido una amplia colección de ejemplos de apropiadas simplificaciones para expresiones numéricas presentadas como porcentajes. Nuestro objetivo es usar estos datos para guiar el desarrollo de un sistema para simplificar automáticamente porcentajes en textos. Nuestras hipótesis

iniciales eran: que el valor de la cantidad original influía en la estrategia de simplificación utilizada y en el uso o no de modificador en la versión simplificada; que la representación matemática de la expresión influía en la accesibilidad de la expresión simplificada y que la pérdida de precisión influía en la estrategia de simplificación y en el uso de modificadores. El análisis de los datos recogidos nos ha permitido comprobar cuáles de estas hipótesis son ciertas a partir de la opinión de los expertos.

4.4.5. Resumen de las estrategias de simplificación de expresiones numéricas identificadas para el inglés

A partir de los datos ya comentados, y otros observados en la encuesta, hemos identificado las estrategias de simplificación de expresiones numéricas para el inglés. A continuación presentamos las principales conclusiones que hemos obtenido:

1. Las expresiones numéricas representadas en letras son transformadas por su correspondiente representación usando dígitos.

2. Independientemente del nivel de dificultad para el que se está simplificando, las estrategias comunes identificadas son:

   - Los porcentajes se redondean al valor más próximo, tanto para valores comunes y no comunes, como para valores centrales y periféricos.
   - Las expresiones no numéricas se usan sólo para los valores extremos de la cantidad.
   - Las fracciones más comunes se usan en el rango central de cantidades, mientras que en los rangos extremos se usan otro tipo de estrategias.

3. Teniendo en cuenta el nivel de dificultad para el que se está simplificando:

   - Si el nivel de dificultad se corresponde con el de una persona que no entiende porcentajes, la estrategia a usar es cambiar las expresiones en porcentajes por sus correspondientes fracciones equivalentes.
   - Si el nivel de dificultad se corresponde con el de una persona que no entiende los porcentajes con decimales, la estrategia a usar es redondear la cantidad al porcentaje más próximo sin decimales.
   - Si el nivel de dificultad se corresponde con una persona que tiene dificultades con las expresiones numéricas de manera más general, las estrategias más usadas son fracciones, seguidas de ratios.

Adaptando las expresiones originales a este tipo de representación matemática se simplifica su dificultad.

4. Con los datos recogidos no observamos un comportamiento claro a la hora de utilizar los ratios como estrategia de simplificación, ni en relación a valores comunes o no comunes, ni en los rangos centrales ni periféricos de la proporción.

5. Independientemente del nivel de dificultad para el que se está simplificando, el uso de modificador no está influenciado ni por el valor central o extremo de la proporción, ni por la estrategia de simplificación utilizada.

6. Si no hay pérdida de precisión en la simplificación, entonces no se usa modificador.

7. No hemos encontrado correlación entre la pérdida de precisión y el uso de modificadores.

8. Hemos observado que si la expresión original tiene modificador, entonces en la expresión simplificada se mantiene.

4.5. Identificación experimental con expertos de las estrategias de simplificación de expresiones numéricas con y sin contexto en español

Para el caso de la identificación de las estrategias de simplificación de expresiones numéricas en español, distintas metodologías de diseño del estudio se han explorado dentro del procedimiento general a seguir propuesto en este trabajo. En todos los casos el objetivo del estudio es esbozar conclusiones sobre el tipo de operaciones de simplificación que podrían ser aplicadas automáticamente a las expresiones numéricas en español.

En este caso, tres metodologías distintas son aplicadas a la hora de diseñar el estudio. En la primera se analiza un corpus paralelo de textos originales en español y su correspondiente versión simplificada a mano. En la segunda se lleva a cabo una encuesta con expertos para identificar las simplificaciones de las expresiones numéricas que ellos realizan. En la tercera, realizamos un estudio con un colectivo concreto, en nuestro caso personas con dislexia.

Dentro de la variedad de tipos posibles de 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*).

A continuación describimos las intuiciones de trabajo planteadas, la selección del material utilizado, el diseño de los estudios y el análisis de los datos recogidos en cada caso.

Los casos del análisis del corpus paralelo y la encuesta con expertos se contemplan a la vez, ya que se realizó un estudio comparativo y se discuten los resultados obtenidos en ambos casos. El caso con usuarios reales lo presentamos como un caso aparte donde el procedimiento a seguir es el mismo, pero partimos de un conjunto de hipótesis distintas y la metodología aplicada difiere un poco de los casos anteriores, de ahí que se contemple como un caso separado.

4.5.1. Intuiciones planteadas para la simplificación de expresiones numéricas en español

Partimos de un conjunto de intuiciones que planteamos en el proceso de simplificación de expresiones numéricas en español, para validarlas con los resultados que obtengamos de los estudios realizados:

1. Las expresiones originales expresadas en letras deben ser simplificadas sustituyéndolas por su correspondiente versión en dígitos.

2. En el proceso de simplificación de la expresión numérica, si hay pérdida de precisión, se añade un modificador y se redondea la cantidad original de la expresión.

3. Si en la expresión numérica original existe modificador, pero hay pérdida de precisión en el proceso de simplificación, se cambia el modificador original y se redondea la cantidad para generar la versión simplificada de la expresión.

4. Las expresiones numéricas originales se reescriben cambiando su representación matemática, por ejemplo de porcentajes a fracciones o de fracciones a ratios.

4.5.2. Selección del material utilizado para la simplificación de expresiones numéricas en español

Se utiliza el corpus recopilado en el proyecto *Simplext*[2], presentado como recurso en el capítulo 3 (sección 3.1). Este corpus paralelo contiene 40 textos con un total de 570 oraciones, 246 oraciones en el conjunto original y 324 oraciones en el conjunto simplificado. Dicho corpus sirvió para documentar todas las operaciones de simplificación aplicadas por los humanos para planificar y organizar su implementación automática. Entre la variedad de

[2]www.simplext.es

operaciones detectadas en este trabajo nos centramos en simplificaciones léxicas, y más específicamente en el tratamiento de las expresiones numéricas, que es el trabajo que presentamos en esta tesis.

Un subconjunto de los textos originales se utilizó para determinar las frases que iban a ser utilizadas en la encuesta que se pasó a los expertos. Se seleccionaron distintas oraciones en las que se podían ver diferentes tipos de expresiones numéricas para ampliar el rango de las transformaciones que aplicaban los humanos a la hora de simplificar las expresiones numéricas originales.

4.5.3. Diseño del estudio para la simplificación de expresiones numéricas en español

Nuestro estudio consta de dos partes. Una es el análisis del corpus paralelo de textos originales y simplificados, y la otra consiste en el diseño, implementación y análisis de una encuesta complementaria similar a la que realizamos para el inglés, con el fin de ampliar nuestro conocimiento sobre posibles simplificaciones de información numérica. Las expresiones numéricas del corpus han sido etiquetadas y extraídas, junto con el resto de la frase donde aparecen, para presentarlas de manera separada en dicha encuesta.

Por lo tanto, por un lado tenemos expresiones numéricas en contexto en el corpus, donde se pueden observar otras operaciones de simplificación, como por ejemplo sustituciones basadas en sinonimia o reestructuración sintáctica. 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.

Se usaron textos del corpus paralelo *Simplext* que fueron 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 pasos a seguir 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.

Los detalles de la identificación y anotación de las expresiones numéricas se presenta junto con el sistema en el capítulo 5, en la sección 5.3.2.

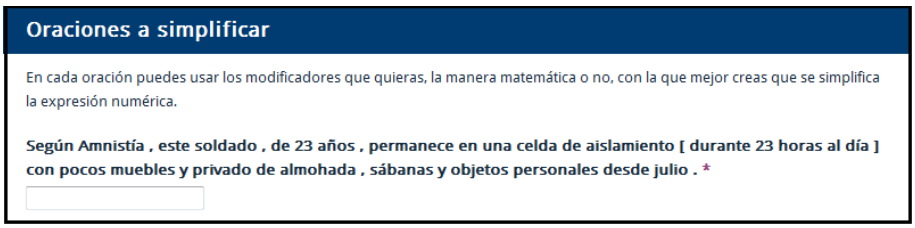


Figura 4.4: Ejemplo de un parte de la encuesta de simplificación de expresiones numéricas en español

La encuesta se implementó utilizando la herramienta que proporciona Google para hacer formularios, *Google Form*, y se albergó en Google Docs[3]. Participaron 23 personas, todos hablantes nativos de español en posesión de un título universitario, diferentes a los que participaron en la edición del corpus paralelo.

El cuestionario se componía de frases tomadas del corpus, con la diferencia de que el contexto que las rodea fue omitido. 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 en la encuesta 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 dado que la idea era compararlas con las operaciones extraídas del corpus y estudiar dicha comparación. La Figura 4.4 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 cual se pedía simplificar.

[3] http://bit.ly/1rZzaDH

4.5.4. Análisis de los datos para la simplificación de expresiones numéricas en español

Aquí presentamos los resultados obtenidos, 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. A continuación, los datos obtenidos se analizan con un enfoque comparativo, con el objetivo de extraer conclusiones para la implementación de reglas generales de simplificación. El estudio del corpus y de la encuesta fueron presentados en el trabajo de Bautista et al. (2012).

4.5.4.1. Análisis del corpus

Como ya se ha mencionado, en este análisis tratamos las 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 necesitan con frecuencia ser modificadas por los expertos para conseguir un texto de salida más fácil de leer. Cada texto original contiene un promedio de 3,78 expresiones numéricas.

En las versiones simplificadas de los textos, un número significativo de estas expresiones numéricas son eliminadas. 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. En trabajos previos (Bautista et al., 2011b; Power y Williams, 2012) ya se sugiere que los modificadores pueden ser una herramienta útil para simplificar una variedad de diferentes expresiones numéricas.

A continuación, presentamos 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 trescientos mil niños[4]

4. Grandes números se redondean:

   casi 7.400 millones de euros ⇒ más de 7000 millones de euros[5]

5. Se aplica redondeo eliminando puntos decimales:

   1,9 millones de hogares ⇒ 2 millones de casas[6]

En la siguiente sección presentamos los datos recogidos en la encuesta que complementa los datos obtenidos a partir del corpus.

4.5.4.2. Resultados de la encuesta

La encuesta está dirigida exclusivamente a la simplificación de expresiones numéricas para analizar el uso de modificadores y las estrategias de simplificación aplicadas. Recopilando esta información, podemos completar los resultados obtenidos en el estudio del corpus antes mencionado de cara a la implementación del sistema de simplificación automática.

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 de la encuesta. Algunos ejemplos son añadir una explicación, calcular el tanto por ciento dado o cambiar de porcentaje a fracción. 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.

Veamos el ejemplo de la expresión original *55* en la siguiente frase, y las simplificaciones que sugirieron los participantes de la encuesta:

> *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 propuestas de simplificación recogidas en la encuesta son las siguientes:

- *más de 50*

[4] Aquí además se redondea la cantidad
[5] Aquí se cambia el modificador
[6] Aquí otro cambio léxico es aplicado: hogar ⇒ casa

- *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 conclusiones generales que sacamos del análisis de datos obtenidos del cuestionario son las siguientes:

- Cuando se simplifica un número, se dan las siguientes opciones:
  - se deja sin cambios *(58.7 %)*
  - 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 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 fue 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. Es el caso de la expresión *durante 23 horas al día*, que fue reescrita por *casi todo el día*.

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 fue igual a la original, es decir, no hubo ningún cambio); y en 2 casos la operación más común fue reescribir completamente 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 la 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*.

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.

4.5.4.3. Análisis comparativo de los resultados obtenidos

Para llevar a cabo un análisis comparativo de los resultados obtenidos en los estudios realizados 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. Las Tablas 4.9 y 4.10 presentan los resultados. Las filas marcadas corresponden a las operaciones que coinciden en ambos casos.

En los resultados obtenidos del análisis del corpus, más del 50 % de las expresiones numéricas fueron eliminadas (considerando las dos operaciones de eliminación), 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*.

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 Expresión* y *Reescribir Expresión*. Obviando los casos

Operaciones de simplificación	**Número de Expresiones**	**% Uso**
Eliminar Expresión	12	44,4 %
Eliminar Oración	7	25,9 %
Misma Expresión	2	7,4 %
Cambiar Modificador + Redondeo	2	7,4 %
Eliminar Modificador + Redondeo	2	7,4 %
Reescribir Expresión	1	3,7 %
Eliminar Modificador + Mismo número	1	3,7 %
Total	**27**	100 %

Tabla 4.9: Operaciones de simplificación obtenidas del análisis del corpus

de eliminación del corpus, son las tres 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 bastante 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 Expresión*, 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 piensa en 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 siempre 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 el contexto de la expresión numérica dentro de la oración también había sido

Operación de simplificación	**Número de Expresiones**	**% Uso**
Añadir Modificador + Redondeo	9	33,3 %
Cambiar Modificador + Redondeo	6	22,2 %
Misma Expresión	5	18,5 %
Reescribir Expresión	5	18,5 %
Mantener Modificador + Redondeo	2	7,4 %
Total	**27**	100 %

Tabla 4.10: Operaciones de simplificación obtenidas del análisis de la encuesta

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 4.5.4.2, 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.

4.5.4.4. Validación con expertos de nuestras intuiciones en el estudio en español

Hemos querido contrastar nuestras intuiciones iniciales (sección 4.5.1) con los datos recogidos en nuestro estudio, tanto de los resultados obtenidos del análisis del corpus como de las respuestas de la encuesta realizada a los expertos.

Para nuestra primera intuición (las expresiones originales expresadas en letras deben ser simplificadas sustituyéndolas por su correspondiente versión en dígitos), podemos observar que es una operación que se contempla en el corpus paralelo, en cambio los participantes de la encuesta no consideraron este tipo de transformaciones.

La segunda intuición (en el proceso de simplificación de la expresión numérica, si hay pérdida de precisión, se añade un modificador y se redondea la cantidad original de la expresión), se corresponde con la operación más

aplicada en la encuesta (con un 33,3 % de uso). En cambio, en el caso del corpus paralelo no se contempla el caso de añadir modificador como una transformación aplicada.

El caso de nuestra tercera intuición planteada (si en la expresión numérica original existe modificador, pero hay pérdida de precisión en el proceso de simplificación, se cambia el modificador original y se redondea la cantidad para generar la versión simplificada de la expresión), es una transformación considerada en ambos casos, tanto para el caso del corpus paralelo (con un 7,4 % de uso) como para el caso de la encuesta (con un 22,2 % de uso). Por lo que tanto en cualquier caso en el que estemos simplificando con o sin contexto, esta operación es importante aplicarla en el proceso de simplificación de expresiones numéricas.

Respecto a la última intuición (las expresiones numéricas originales se reescriben cambiando su representación matemática), en los datos recogidos y analizados de la encuesta se consideran diferentes casos de reescritura, con cambios en su representación matemática como en los cambios de porcentaje a ratio, o de porcentajes a fracciones. Esto ocurre con una frecuencia de uso de un 18,5 %. En cambio, en el análisis del corpus no se ha contemplado como una operación de simplificación aplicada para generar la versión simplificada.

4.5.4.5. Discusión del análisis de los datos para el español

Después de haber recogido los datos y de haber realizado el análisis de los mismos, a partir del corpus y de la encuesta, en esta sección presentamos unas líneas de discusión general.

En primer lugar, señalar que los casos de eliminación, de la oración entera o 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 ello, en las oraciones de la encuesta no se producen casos de eliminación de la expresión numérica o de la oración completa, ya que no se daba información añadida de dónde aparecía la oración en el texto original.

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). Esto demuestra que hay operaciones que, a priori, son independientes del contexto, y que se aplican en ambos casos, obteniendo una versión simplificada de la expresión numérica que se quiere adaptar.

El estudio realizado en esta tesis para la simplificación de expresiones numéricas en español, apoya las conclusiones previas de los trabajos de (Bautista et al., 2011b) y (Power y Williams, 2012), sobre el uso de modificadores

y el uso de distintas estrategias de simplificación.

4.5.5. Resumen de las estrategias de simplificación de expresiones numéricas identificadas para el español

Hemos llevado a cabo un diseño del estudio a partir de un corpus paralelo de textos originales y sus correspondientes versiones simplificadas a mano, junto con una encuesta con expertos a los que les pedimos que simplificaran las expresiones numéricas presentes en las oraciones que se les mostraban. A partir de ambos estudios empíricos, las estrategias de simplificación de expresiones numéricas identificadas para el español son las siguientes:

1. En la simplificación con contexto hemos observado que:

   - La información numérica entre paréntesis se elimina.
   - En muchos casos, la información numérica directamente se elimina en lugar de intentar simplificarla.
   - Las expresiones representadas en letras se cambian por expresiones representadas en dígitos.
   - Usando información del contexto, a veces las expresiones son reescritas completamente con expresiones no numéricas.
   - Se redondea la cantidad de la expresión y a veces se añade o se cambia el modificador para subsanar la pérdida de precisión.

2. En la simplificación sin contexto hemos observado que:

   - Si la expresión original tiene modificador, casi siempre éste se cambia y se redondea la cantidad. A veces se deja el mismo modificador.
   - Hay muy pocos casos de reescritura, ya que no se accede al contexto.
   - Si la expresión original no tiene modificador, entonces se añade modificador y se redondea la cantidad.

Para ambos casos se ha observado que las expresiones numéricas a veces no se modifican. Esto suele ocurrir en los casos en los que la expresión numérica ya es simple por sí misma, porque es un valor frecuente o un valor ya redondeado, como por ejemplo, 1/4 ó 50 %.

4.6. Identificación experimental de las estrategias de simplificación de expresiones numéricas en español con personas con dislexia

Otra metodología considerada en nuestro trabajo es realizar la identificación experimental de las estrategias de simplificación de expresiones numéricas en español con usuarios reales. En nuestro caso diseñamos el estudio teniendo en mente a un colectivo concreto como son las personas con dislexia. Queríamos comprobar para este colectivo:

- Si los números redondeados son más fáciles de leer y entender que los valores exactos.
- Si la *legibilidad* y *comprensión* eran diferentes al utilizar fracciones o porcentajes.
- Si los números expresados en dígitos eran más legibles y comprensibles que los números expresados en letras.

Antes de presentar cada parte del estudio, vamos a definir dos términos que utilizamos durante nuestro estudio:

- *Legibilidad*: cualidad que indica que la representación de la información puede ser fácilmente leída.
- *Comprensión*: capacidad o facultad para entender lo que se está leyendo.

El objetivo de nuestro estudio fue estudiar cómo la representación numérica afecta a la legibilidad y comprensión de un texto para nativos españoles con y sin dislexia. En las siguientes secciones presentamos las intuiciones planteadas en el estudio, el material utilizado, el diseño del estudio y el análisis y discusión de los datos recogidos. Este trabajo se realizó en colaboración con la Doctora Luz Rello, del grupo TALN de la Universitat Pompeu Fabra de Barcelona. Más detalle sobre este trabajo se puede ver en Rello et al. (2013).

4.6.1. Intuiciones planteadas para las personas con dislexia

Para realizar el estudio con personas con dislexia, formulamos nuestras intuiciones en forma de hipótesis de trabajo que queríamos explorar. A continuación se presentan las hipótesis concretas para los experimentos:

HD1.1: La *legibilidad* del texto mejorará si las expresiones numéricas están representadas en dígitos en lugar de en letras. (*20* vs. *veinte*)

HD1.2: La *comprensión* del texto mejorará si las expresiones numéricas están representadas en dígitos en lugar de en letras. (*20* vs. *veinte*)

HD2.1: La *legibilidad* del texto mejorará si las expresiones numéricas usadas están redondeadas en lugar de usar valores no redondeados, con decimales. (*48* vs. *48,3*)

HD2.2: La *comprensión* del texto mejorará si las expresiones numéricas usadas están redondeadas en lugar de usar valores no redondeados, con decimales. (*48* vs. *48,3*)

HD3.1: La *legibilidad* del texto mejorará si las expresiones numéricas están representadas en porcentajes en lugar de en fracciones. (*25 %* vs. *1/4*)

HD3.2: La *comprensión* del texto mejorará si las expresiones numéricas están representadas en porcentajes en lugar de en fracciones. (*25 %* vs. *1/4*)

4.6.2. Selección del material utilizado

El material utilizado en el estudio es un conjunto de textos creados por Luz Rello y la autora de esta tesis, teniendo en cuenta características como longitud, número de expresiones numéricas, número de entidades nombradas, etc.

Cada participante tenía que leer varios textos en español con diferentes representaciones de expresiones numéricas. Los textos utilizados en los experimentos tienen una longitud media de 62,33 palabras, tienen el mismo número de expresiones numéricas para cada par de textos utilizados en los experimentos, son de temática similar, contienen el mismo número de oraciones y no contienen entidades nombradas, palabras de otros idiomas ni acrónimos.

La Tabla 4.11 muestra un ejemplo de un texto de los utilizados en los experimentos. Hablamos de pares de textos, porque habían textos con expresiones en letras y expresiones en dígitos, textos con expresiones con valores exactos y con valores redondeados y textos con expresiones en fracciones y con expresiones en porcentajes, para poder hacer las distintas comparativas.

4.6.3. Diseño del estudio con personas con dislexia

Para realizar el estudio preparamos un experimento usando un *eye-tracker* (un aparato que hace seguimiento de los ojos a la hora de leer en la pantalla de un ordenador) para recoger y analizar los datos que nos permitieran contestar a nuestras preguntas.

Con esto en mente, condujimos tres experimentos con 72 personas (36 de ellas con dislexia) usando un *eye-tracker* y encuestas de comprensión para analizar el proceso de lectura y la comprensión del texto.

Composición de una hamburguesa
El pan supone entre ***el 30 %*** *y*
el 50 % *del peso de una*
hamburguesa. La hamburguesa tiene
un valor energético que oscila entre
las ***250*** *y* ***300 kilocalorías****. Un adulto*
con actividad moderada necesita en torno
a ***2.500 kilocalorías*** *diarias, por lo que*
una hamburguesa a la semana no desequilibra
ninguna dieta ni siquiera incorporándole un
sobre de ***11 gramos*** *de ketchup,*
que contiene ***70 kilocalorías.***

Tabla 4.11: Ejemplo de uno de los textos utilizados en uno de los experimentos con personas con dislexia.

Hay que tener en cuenta que cuando leemos un texto el ojo no se mueve de forma contigua sobre el texto, sino que alterna movimientos sacádicos y fijaciones visuales, es decir, avanza dando saltos cortos y se detiene en diferentes partes del texto. La duración de la fijación denota el tiempo que el ojo descansa todavía en un solo lugar del texto. Se ha demostrado que esta medida es un indicador válido de la *legibilidad* de un texto. De acuerdo con diversos trabajos de investigación, (Just y Carpenter, 1980; Rayner y Duffy, 1986; Sereno y Rayner, 2003) las fijaciones de duración más corta están asociadas con una mejor lectura, mientras que las fijaciones más largas pueden indicar mayores cargas de procesamiento del texto. Por lo tanto, utilizamos la duración de las fijaciones como una medida para cuantificar la lectura del texto. Para medir la *comprensión* del texto leído, usaremos cuestionarios, uno por texto. Definimos varias preguntas de opción múltiple, con tres posibles opciones, una correcta y dos incorrectas. Para estas respuestas, hemos calculado el porcentaje de respuestas correctas, donde cada opción correcta está puntuada con 100 % de acierto y las otras con 0 %.

En la Figura 4.5 podemos ver dos ejemplos de las preguntas, de información inferencial (i) y de información detallada (ii), que se presentaron a los sujetos.

4.6.4. Análisis de los datos recogidos

La validación de nuestra hipótesis por parte de los participantes se hizo, por un lado a partir de los datos recogidos por el *eye-tracking*, y por otro lado a partir de los datos recogidos en la encuesta que se les pasó después de cada experimento. Las medidas usadas para la comparación de los textos fueron las medias de la duración de la fijación, la duración total de lectura y las respuestas correctas. Consideramos el grupo N como el grupo de parti-

(i) El texto trata sobre:

(a) La descomposición de una hamburguesa.

(b) La creación de una hamburguesa.

(c) La composición de una hamburguesa.

(ii) Una porción de patatas fritas tiene un máximo de:

(a) 200 kilocalorías.

(b) 300 kilocalorías.

(c) 400 kilocalorías.

Figura 4.5: Dos ejemplos de las preguntas de los cuestionarios de comprensión del experimento con personas con dislexia

cipantes no disléxicos, mientras que el grupo D se corresponde con el grupo de personas disléxicas que formaron parte de los experimentos. Para comprobar nuestras hipótesis, las diferencias entre grupos y condiciones fueron comprobadas con las medias de los tests de la *t-Student*.

Para la hipótesis HD1.1 (la *legibilidad* del texto mejorará si las expresiones numéricas están representadas en dígitos en lugar de en letras), no encontramos significancia estadística en la *legibilidad* para el grupo N ($p < 0{,}444$) aunque el tiempo de fijación disminuye cuando leen expresiones numéricas en dígitos. Sin embargo, encontramos significancia estadística para la *legibilidad* en el grupo D teniendo en cuenta la media del tiempo de fijación ($p < 0{,}054$). Este resultado apoya nuestra hipótesis. En cuanto a la hipótesis HD1.2 (la *comprensión* del texto mejorará si las expresiones numéricas están representadas en dígitos en lugar de en letras), aunque el uso de dígitos incrementa el número de respuestas correctas que reflejan una mayor *comprensión* del texto, no hay significancia estadística para ambos grupos. Por tanto, se rechaza la hipótesis ($p < 0{,}241$ para el grupo N y $p < 0{,}269$ para el grupo D).

También se rechaza la hipótesis HD2.1 (la *legibilidad* del texto mejorará si las expresiones numéricas usadas están redondeadas en lugar de usar valores no redondeados, con decimales), ya que no encontramos significancia estadística en la *legibilidad* del texto en el grupo N ($p < 0{,}867$) ni en el grupo D ($p < 0{,}685$) cuando leen texto con expresiones numéricas redondeadas, teniendo en cuenta la media del tiempo de fijación. En ambos casos, el tiempo de fijación se incrementa porque en las expresiones redondeadas los participantes tienen que leer más palabras (modificador y cantidad) que en el texto con las expresiones numéricas con decimales. Sin embargo, para la hipótesis HD2.2 (la *comprensión* del texto mejorará si las expresiones numéricas usadas están redondeadas en lugar de usar valores no redondeados, con decimales), las expresiones numéricas redondeadas han ayudado a nuestro participantes no disléxicos y su tasa de respuestas correctas se incrementa para las cuestiones inferenciales, pero no para los participantes con

dislexia. Sin embargo, aunque el número de respuestas correctas crece, estos resultados no soportan nuestra hipótesis porque no encontramos significancia estadística para la *comprensión* en ambos grupos ($p < 0{,}310$ en el grupo N y $p < 0{,}695$ en el grupo D).

En el último experimento no encontramos significancia estadística en la *legibilidad* del texto para el grupo N ($p < 0{,}462$) teniendo en cuenta la media del tiempo de fijación. Sin embargo, nuestros resultados confirman nuestra hipótesis HD3.1 (la *legibilidad* del texto mejorará si las expresiones numéricas están representadas en porcentajes en lugar de en fracciones) porque encontramos significancia estadística para la *legibilidad* del texto en el grupo D ($p < 0{,}046$) cuando leen textos con expresiones numéricas en porcentajes. Este grupo lee más rápido los textos con expresiones numéricas en porcentajes que textos con información numérica en fracciones. Por otro lado, rechazamos la hipótesis HD3.2 (la *comprensión* del texto mejorará si las expresiones numéricas están representadas en porcentajes en lugar de en fracciones) porque no encontramos significancia estadística en los resultados para la *comprensión* del texto en ambos grupos ($p < 0{,}170$ para el grupo N y $p < 0{,}474$ para el grupo D).

Cada participante rellenó además un cuestionario con 20 preguntas que fueron evaluadas usando una escala de Likert de 5 niveles. Para 10 de las preguntas a los participantes se les preguntó sobre cómo de fácil fue leer el texto (*legibilidad*); mientras que con las otras 10 preguntas a los participantes se les preguntó cómo de fácil fue comprender el texto (*comprensión*). Los participantes encontraron significativamente más fáciles de leer los números escritos en dígitos que los números escritos en letras ($p < 0{,}001$) además de más comprensibles ($p < 0{,}001$). Pero no encontramos significancia estadística para la lectura entre los números redondeados o sin redondear ($p = 0{,}272$) ni en la comprensión del texto ($p = 0{,}446$). Finalmente, los participantes encontraron significativamente más fáciles de leer los porcentajes que las fracciones ($p < 0{,}001$) además de más comprensibles ($p < 0{,}001$).

4.6.4.1. Discusión de los datos recogidos en el estudio con personas con dislexia

Con respecto al estudio llevado a cabo con personas con dislexia, podemos ver que las diferencias entre el uso de dígitos y el uso de expresiones numéricas en letras indican una mejora estadísticamente significativa para la lectura del texto en personas con dislexia cuando se utilizan dígitos. Esto se debe a que las expresiones numéricas descritas usando letras requieren un número más largo de palabras en comparación con su correspondiente versión usando dígitos. En general, la longitud es un parámetro ya conocido que crea dificultades a las personas con dislexia, de ahí que la reducción de longitud que conlleva la utilización de dígitos debería siempre hacer más fácil la lectura a este colectivo.

Con respecto a las diferencias entre el uso de números redondeados o sin redondear, a la hora de la legibilidad del texto no encontramos diferencias estadísticamente significativas. Aún así, ambos grupos aumentan su tiempo de lectura del texto cuando se usan números redondeados. Esto puede ser debido al aumento de la longitud de las expresiones, ya que en el proceso de redondeo se añaden modificadores como *casi* o *más de* para indicar la pérdida de precisión. Esto también puede explicar los resultados para comprensión del texto que hemos observado en el grupo de disléxicos, ya que tienen que leer y comprender más palabras cuando se redondea la expresión. Sin embargo, en el grupo de no disléxicos hay un incremento, no significativo, de la comprensión del texto cuando se usan números redondeados, lo cual cuadra con nuestra hipótesis de que los números redondeados son más fáciles de entender a pesar de contener menos detalles y perder precisión matemática.

Con respecto a las diferencias entre el uso de porcentajes y fracciones, hay un incremento estadísticamente significativo en la legibilidad del texto para el grupo de disléxicos cuando se usan porcentajes en lugar de fracciones. Por el contrario, hay un descenso en la legibilidad del texto para el grupo de no disléxicos en la misma situación. De nuevo, los resultados para comprensión del texto no son estadísticamente significativos en ninguno de los dos casos. Sin embargo, los resultados obtenidos por ambos grupos indican que la comprensión del texto disminuye cuando se usan porcentajes en lugar de fracciones.

Hay una aparente contradicción en el grupo de disléxicos, para los cuales los porcentajes parecen ser más fáciles de leer pero más difíciles de entender. Una posible explicación puede estar relacionada con la naturaleza de estas expresiones. Desde un punto de vista conceptual, tanto porcentajes como fracciones expresan una proporción relativa entre dos cantidades: el valor del porcentaje y 100 en el caso de los porcentajes, y el valor del numerador y el valor del denominador en el caso de las fracciones. Sin embargo, el valor al que se hace referencia en el caso de los porcentajes es implícito (producido por el signo %). Esto implica que para las fracciones, dos cantidades tienen que leerse, mientras solo una cantidad tiene que leerse para los porcentajes. Esto puede explicar la relativa facilidad para el grupo de disléxicos a la hora de leer porcentajes (sólo una cantidad para leer) frente a las fracciones (dos cantidades diferentes para leer). Por el contrario, el hecho de que los dos valores comparados son explícitos en la expresión para el caso de la fracciones puede que las hagan más fáciles de entender que los porcentajes, donde el valor de referencia tiene que ser inferido.

4.6.5. Resumen de las estrategias de simplificación de expresiones numéricas identificadas en español para personas con dislexia

Una vez realizado el estudio con personas con dislexia, a partir de las hipótesis de trabajo planteadas y el análisis de los datos recogidos, podemos afirmar que las estrategias de simplificación de expresiones numéricas, que hemos identificado para este colectivo en concreto que son las personas con dislexia, son las siguientes:

1. Las personas con dislexia leen mejor las expresiones numéricas representadas en dígitos.
2. La estrategia de simplificación de redondear la cantidad original y acompañarla de un modificador para generar la versión simplificada de la expresión numérica, aumenta el tiempo de lectura y no mejora la comprensión para las personas con dislexia, ya que tienen que leer más.
3. Las personas con dislexia prefieren leer expresiones numéricas representadas en porcentajes, frente a la representación en fracciones, aunque les resulte más difíciles de comprender, porque tienen que inferir el valor de referencia en el porcentaje.

4.7. Comparación de las estrategias de simplificación de expresiones numéricas identificadas para el inglés y para el español

La identificación experimental de las estrategias de simplificación de expresiones numéricas ha sido llevada a cabo con distintas metodologías para el inglés y para el español. Una vez realizados estos estudios se han identificado un conjunto de estrategias de simplificación. Se ha realizado una comparación entre los dos idiomas estudiados para analizar qué estrategias tienen en común y qué características presentan las estrategias para cada idioma. Estas estrategias son la base de la implementación de los sistemas automáticos que hemos desarrollado para cada idioma y que presentamos en el siguiente capítulo.

Recordemos que las características propias del estudio de simplificación de expresiones numéricas en inglés son:

1. El estudio sólo contempla la simplificación de expresiones numéricas representadas en porcentajes.
2. Contempla el uso de expresiones no numéricas para los rangos extremos de la proporción.

3. Las expresiones numéricas representadas en fracciones son simplificadas utilizando fracciones equivalentes más comunes acompañadas de un modificador.

Recordemos que las características propias del estudio de simplificación de expresiones numéricas en español son:

1. El estudio contempla un rango más amplio de tipos de expresiones numéricas: numerales, monetarias, porcentajes y fracciones.
2. No se hace uso de expresiones no numéricas como candidatas a utilizar en la versión simplificada.
3. Las expresiones representadas en fracciones no se adaptan a versiones equivalentes.

En ambos estudios e independientemente del idioma para el que se quiera simplificar, hemos observado que:

1. Las expresiones representadas en letras deben ser transformadas en su correspondiente versión en dígitos.
2. Las expresiones numéricas en porcentajes se deben redondear al valor más próximo para generar la versión simplificada de la expresión.
3. Los ratios no se modifican en el proceso de simplificación.
4. Si no hay pérdida de precisión en la simplificación, no se usa modificador en la versión simplificada.
5. El uso de modificador en la versión simplificada de la expresión contempla varias opciones:
   - Si existe modificador, se mantiene.
   - Si no existe modificador, se añade.

Además, en el estudio de simplificación de expresiones numéricas es español con personas con dislexia hemos identificado que:

1. Las expresiones representadas en dígitos son mejores de leer para las personas con dislexia.
2. El uso de modificadores acompañando a una cantidad redondeada en una versión simplificada de una expresión numérica, aumenta el tiempo de lectura y no mejora la comprensión para las personas con dislexia, porque tienen que leer más.
3. Las personas con dislexia prefieren leer las expresiones representadas en porcentajes, aunque les resulte más difíciles de comprender porque tienen que inferir el valor de referencia.

Resumen y conclusiones

En este capítulo hemos presentado la descripción y las etapas correspondientes al modelo genérico para la simplificación que utilizamos en la implementación de los sistemas que presentamos en el siguiente capítulo.

Además, hemos presentado distintas metodologías planteadas en el procedimiento a seguir para la identificación de estrategias de simplificación necesarias para automatizar el proceso de simplificación de expresiones numéricas. Hemos realizado distintos estudios en inglés y en español que nos han permitido obtener conclusiones y algunas de ellas son utilizadas por los sistemas implementados como parte de este trabajo.

De los estudios realizados hemos aprendido que a la hora de simplificar expresiones numéricas en porcentajes en inglés, dependiendo del nivel de dificultad para el que se está simplificando, la estrategia que se usa es diferente, de ahí que las fracciones se utilicen cuando una persona no entienda porcentajes y que se redondee al porcentaje más próximo sin decimales cuando la expresión sea un porcentaje con decimales y la persona tenga dificultades con los decimales. En todos los casos, las información numérica debe ser presentada en dígitos. Además, independientemente del nivel de dificultad para el que se está simplificando, los porcentajes se redondean al valor más próximo, las fracciones más comunes se usan en el rango central mientras que en los rangos extremos se usan otro tipo de estrategias, como es el caso de las expresiones no numéricas que se usan mayoritariamente en los rangos extremos. Hemos observado que el uso de modificador no está influenciado ni por el valor de la proporción ni por la estrategia de simplificación utilizada. No hemos encontrado correlación entre la pérdida de precisión y el uso de modificadores, pero si no hay pérdida de precisión en la simplificación entonces no se usa modificador. Además, si la expresión original tiene modificador, éste se mantiene en la expresión simplificada.

Para el caso de la simplificación de expresiones numéricas en español, hemos observado que teniendo en cuenta o no el contexto a la hora de realizar la simplificación pueden considerarse distintas operaciones. Cuando se simplifica con contexto se tiende a eliminar información numérica en general y sobre todo la información entre paréntesis, las expresiones en letras se transforman en dígitos, a veces se reescriben las expresiones y se simplifican redondeando la cantidad y añadiendo o cambiando el modificador. Cuando se simplifica sin contexto, la opción de eliminar no se contempla, se tiende a simplificar las expresiones redondeando la cantidad y cambiando o dejando el modificador original, o se reescriben o se dejan igual. Estas tres operaciones son comunes con la simplificación con contexto.

Del estudio realizado con usuarios reales, en nuestro caso con personas con dislexia, hemos observado que las personas con dislexia leen mejor las expresiones representadas en dígitos. Para este colectivo concreto, el uso de

cantidades redondeadas acompañadas de un modificador aumenta el tiempo de lectura y no mejora la comprensión, ya que tienen que leer más. A la hora de preferir una representación matemática u otra, las personas con dislexia prefieren las expresiones en porcentajes frente a fracciones aunque les sean más difíciles de entender porque tienen que inferir el valor de referencia del porcentaje.

En el siguiente capítulo presentamos el desarrollo e implementación de dos sistemas de simplificación de expresiones numéricas en inglés y en español que siguen el modelo genérico presentado y utilizan las pautas descubiertas en estos casos de estudio de identificación experimental llevados a cabo. Para ambos sistemas se lleva a cabo una evaluación de los mismos con expertos que nos permite evaluar la salida de nuestros sistemas y plantear mejoras en los mismos. El diseño de la evaluación consiste en un cuestionario con un subconjunto de oraciones originales y su correspondiente versión simplificada generada por el sistema. A los participantes del estudio se les pregunta si la versión simplificada que les presenta, según su criterio, es correcta o no para la oración original dada. Los detalles de la evaluación y el análisis de los datos recogidos los presentamos en la sección de evaluación de cada sistema.

Capítulo 5

Sistemas de simplificación de expresiones numéricas

En el capítulo anterior recogimos las bases teóricas para diseñar e implementar los sistemas de simplificación automática de expresiones numéricas. Por un lado presentamos el modelo de proceso que siguen nuestros sistemas y por otro lado recogimos un conjunto de estrategias de simplificación identificadas a partir de distintas metodologías llevadas a cabo.

El objetivo de este capítulo es presentar los sistemas implementados, uno para el inglés y otro para el español. Para cada sistema implementado vamos a ir viendo las decisiones tomadas en cada etapa, las herramientas que hemos usado y las variables que se configuran en cada etapa. Junto con la descripción de cada sistema, presentamos la evaluación que se ha realizado del mismo.

5.1. Sistema de simplificación de expresiones numéricas en inglés

Presentamos el diseño e implementación de nuestro sistema para simplificar de manera automática las expresiones numéricas en textos en inglés. La primera variable que hay que configurar en el modelo específico (sección 4.2) que sigue nuestro sistema es el lenguaje del texto original que se quiere simplificar, que en nuestro caso ya sabemos que es el inglés. Este sistema fue presentado en el trabajo de Bautista et al. (2013b).

En este sistema concreto para textos en inglés, la variable correspondiente al nivel de dificultad al que se quiere adaptar las expresiones numéricas se considera en la etapa donde las operaciones de simplificación son elegidas. A partir de la escala de conceptos definida en la sección 4.4.3, hemos considerado tres niveles diferentes de dificultad en los que puede trabajar el sistema, de menor a mayor dificultad de comprensión de las expresiones numéricas:

1. *Fractions*: nivel correspondiente a cuando el usuario sólo entiende expresiones numéricas en forma de fracción o con dificultad menor.
2. *Percentages without decimals (PWD)*: nivel correspondiente a cuando el usuario sólo entiende expresiones numéricas representadas en porcentajes sin decimales o de dificultad menor según la escala de aprendizaje.
3. *Percentages with decimals*: Este es el nivel más difícil, donde el usuario entiende porcentaje con decimales o expresiones numéricas de dificultad menor.

Usando una aplicación con interfaz gráfica en Java, el usuario puede cargar un texto en el sistema, elegir el nivel de dificultad al que quiere adaptar las expresiones numéricas, de entre los que proporciona el sistema, y obtener un texto de salida con las expresiones numéricas simplificadas. En la Figura 5.1 podemos ver una captura de la aplicación.

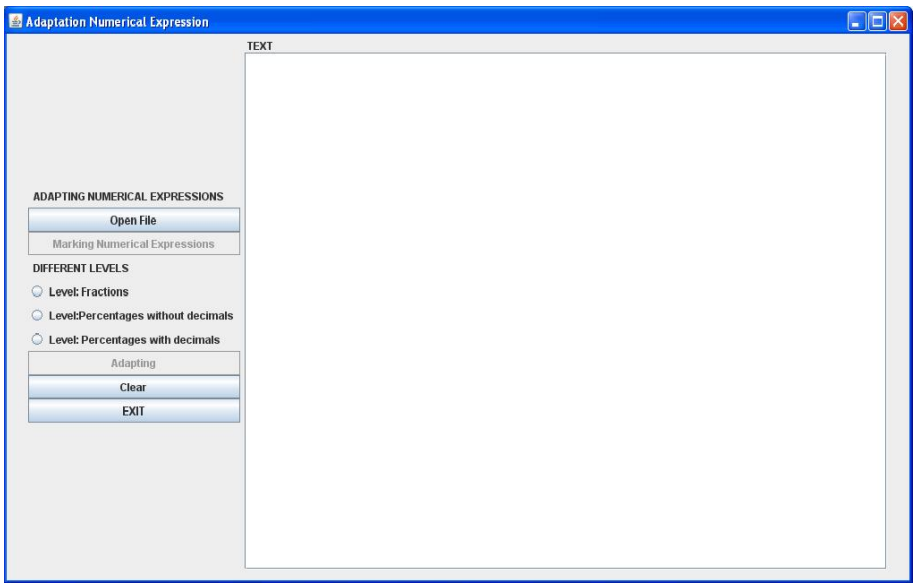


Figura 5.1: Interfaz del sistema desarrollado para el inglés

El sistema funciona sólo con expresiones numéricas en los más altos niveles de la escala (niveles más difíciles). Así, las expresiones numéricas tratadas corresponden a porcentajes o porcentajes con decimales y serán adaptadas a niveles de menor dificultad. En cualquier momento el usuario puede seleccionar el nivel al que adaptar las expresiones numéricas originales del texto usando la interfaz del sistema.

Para demostrar el funcionamiento del sistema usaremos un texto del corpus del proyecto *NumGen* (Williams y Power, 2010). Podemos ver que es un texto rico en expresiones numéricas, formado por 13 oraciones y que contiene 15 expresiones numéricas, muchas de ellas representadas en porcentajes.

Another record year for A-levels
The A-level pass rate rose for the 26th year in a row as record number of teenagers achieved top grades. But figures released by the exam boards highlighted startling discrepancies in Grade A pass rates between regions across England. Statistics from the exam boards showed greater improvements in students in the South East getting A grades in the past **six** years than those in the North East. The South East has seen a **6.1 %** increase in A grades - to **29.1 %** - since 2002 but the North East has seen an improvement of only **2.1 %** - to **19.8 %** - during the same period. But the percentage of pupils gaining passing E grades is rising quicker in the North East - an improvement of **3.4 %** in **six** years compared with **2.8 %** in the South East. Overall figures showed the national pass rate soared **above 97 %** for the first time this year, while **one in four** **sixth**-formers were awarded A grades (**25.9 %**, up from **25.3 %** last year). The figures showed traditional subjects are still firm favourites with English and maths the top choices for candidates. Dr Mike Cresswell, director general of the AQA, said A-levels remained a "highly-valued qualification". He said he was particularly pleased to see the numbers of maths candidates rise from **60,093** last year to **64,593** this year. "There was an upward trend that began a couple of years ago that has accelerated. There are more candidates doing mathematics than at any time in the past. It's important we have people with high mathematic skills so that has to be good news."

Partiendo del modelo de simplificación de expresiones numéricas tenemos que instanciar las variables que entran en juego en el sistema que hemos implementado. El funcionamiento general del sistema es el que sigue. Como el idioma con el trabajamos es el inglés, decidimos usar el parser desarrollado por Williams (2010) (presentado en la sección 3.3.1) que identifica y anota las expresiones numéricas que hay en el texto. El sistema está integrado con el programa de aproximación de proporciones (Power y Williams, 2012) (presentado en la sección 3.3.2) y que usamos en la etapa de simplificación del texto para calcular los candidatos para simplificar cada expresión numérica. Las reglas que definimos a partir de lo aprendido en la identificación experimental realizada están implementadas en Java. Trabajamos con un tipo de texto concreto que son las noticias de prensa. El nivel de dificultad está definido en tres niveles distintos y finalmente la salida del sistema es un texto plano correspondiente al texto original con las expresiones numéricas simplificadas.

En las secciones siguientes se describen las etapas de nuestro modelo y se explica cómo se trabaja en cada etapa y las decisiones que se tienen que tomar. En la Figura 5.2 podemos ver como se instancia para este sistema el modelo de simplificación de expresiones numéricas (presentado en la Figura 4.2).

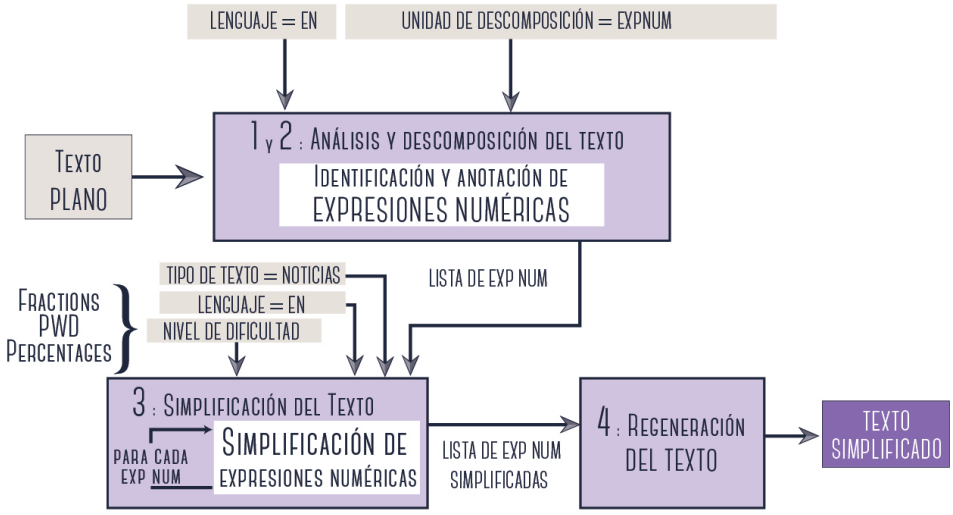


Figura 5.2: Etapas del modelo automático de simplificación centrado en expresiones numéricas para el inglés tal y como se ha instanciado para el sistema de simplificación de expresiones numéricas en inglés

5.1.1. Etapas 1 y 2: Análisis y descomposición del texto

En nuestro sistema, las etapas 1 (Análisis del Texto) y 2 (Descomposición del Texto) del modelo se realizan a la vez. En el propio análisis del texto se identifican y se anotan las expresiones numéricas del texto de entrada.

La etapa de descomposición del texto corresponde a la identificación de las unidades de descomposición que se quieren simplificar, en nuestro caso, las expresiones numéricas identificadas y anotadas. La salida de esta etapa es una lista de expresiones numéricas.

En la siguiente sección vamos a ver con detalle cómo se realizan la identificación y anotación de las expresiones numéricas en los textos en inglés.

5.1.1.1. Etapa 2.1 y 2.2: Identificación y anotación de expresiones numéricas

Para el texto elegido, el sistema usa el *parser* específicamente desarrollado por Williams (2010) para identificar y anotar las expresiones numéricas que hay en el texto. El resultado del análisis se guarda en un archivo XML. El *parser* utiliza reglas basadas en patrones para identificar cinco tipos diferentes de expresiones numéricas: porcentajes, fracciones, cardinales, decimales y expresiones monetarias. Además, clasifica los modificadores en cuatro tipos: *greaterthan*, *lessthan*, *approx* y *exact*. La clasificación de los modificadores se hace según un listado de modificadores con el que cuenta el analizador, según

su significado lingüístico. Si el modificador indica *por encima de* se clasifica según el primer tipo *greaterthan*, si indica *por debajo de* se clasifica según el segundo tipo *lessthan*, si indica un valor *aproximado* se clasifica según el tercer tipo *approx* y si indica un valor *exacto* se clasifica según el cuarto tipo *exact*.

Cada expresión numérica en el texto se identifica con una etiqueta "*<numex>*" en el archivo XML resultante. Diferentes atributos se añaden a esta etiqueta para anotar las diferentes características de la expresión identificada. Entre estos atributos se incluye el modificador que acompaña a la cantidad numérica de la expresión, si lo hay. Por ejemplo, en la oración *Overall figures showed the national pass rate soared above 97 % for the first time this year...* se identifica la expresión numérica *above 97 %* con su modificador y su cantidad expresada en porcentaje.

Para completar la identificación hecha, se anotan una serie de atributos que son añadidos a las etiquetas de las expresiones numéricas durante el análisis del *parser*. Estos atributos guardan las características de cada expresión numérica identificada. Los atributos son:

1. **type**: anota el tipo de expresión numérica (cardinal, fracción, porcentaje o porcentaje decimal).

2. **format**: anota el formato de la expresión (dígitos o letras).

3. **Vg (*given value*)**: es el valor de la cantidad de la expresión. Las expresiones se normalizan entre 0 y 1 si son porcentajes.

4. **units**: anota las unidades, si existen en la expresión.

5. **hedge**: anota el modificador, si existe.

6. **hedge-sem**: clasifica el modificador, según los cuatro tipos que tiene definido el *parser* (*greaterthan*, *lessthan*, *approx* y *exact*).

La salida del *parser* es un documento XML donde cada oración es identificada y cada expresión es anotada. En el siguiente fragmento de código XML podemos ver un ejemplo de como una expresión numérica es anotada por el *parser*:

```
<doctype w3c doctype="numgen">
    ...
    Overall figures showed the national pass rate soared
    <numex hedge="above" hedge-sem="greaterthan"
        type="percentage" format="digits" Vg="0.97">
        above 97 %
    </numex>
    ...
</doctype>
```

A partir del archivo XML producido por el *parser*, obtenemos la siguiente información para la expresión numérica *above 97 %* del ejemplo del fragmento anterior:

1. **type**: *percentage*. Indica que la expresión numérica es un porcentaje.
2. **format**: *digits*. Indica que está expresada en dígitos.
3. **Vg (*given value*)**: Indica el valor de la proporción, Vg=0.97. En este caso el valor está normalizado al ser la expresión numérica un porcentaje.
4. **units**: en la expresión analizada no hay unidades.
5. **hedge**: *above*. Indica el modificador de la expresión, en este caso *above*.
6. **hedge-sem**: *greaterthan*. Indica el tipo de modificador según su significado y la clasificación del *parser*. *Above* indica *por encima de*, por eso se clasifica en el primer tipo.

Si, por ejemplo, tenemos la expresión numérica *25.9 %* en el texto, tendríamos la siguiente información en la anotación:

1. **type**: *percentage*. Indica que la expresión numérica es un porcentaje.
2. **format**: *digits*. Indica que está expresada en dígitos.
3. **Vg (*given value*)**: Indica el valor dado de la proporción, Vg=0.259. En este caso el valor está normalizado al ser la expresión numérica un porcentaje.
4. **units**: en la expresión analizada no hay unidades.
5. **hedge**: no hay modificador.
6. **hedge-sem**: tampoco hay tipo.

Finalmente, si tuviéramos una expresión como *around twenty grams*, obtendríamos la siguiente información:

1. **type**: *cardinal*. Indica que la expresión numérica es un número cardinal.
2. **format**: *letters*. Indica que está expresada en letras.
3. **Vg (*given value*)**: Indica el valor de la expresión, Vg=20. Como no es un porcentaje, no se normaliza.
4. **units**: *grams*. Indica que las unidades son gramos.
5. **hedge**: *around*. Indica el modificador de la expresión.
6. **hedge-sem**: *approx*. Indica el tipo de modificador según su significado y la clasificación del *parser*. *Around* indica un valor *aproximado*, por eso se clasifica según el tercer tipo.

5.1.2. Etapa 3: Simplificación del texto

En esta etapa, las reglas de simplificación específicas para cada caso dependen del nivel de dificultad elegido por el usuario en el sistema. Así, el sistema tiene que adaptar cada expresión numérica identificada y anotada en la etapa anterior para generar la versión simplificada.

El proceso de simplificación tiene dos fases. En una primera fase se utiliza un programa que obtiene una lista de candidatos a utilizar como posibles opciones simplificadas de la expresión numérica original. De ellos, la aplicación elige el mejor candidato. En la segunda fase se aplican las reglas de simplificación que hemos definido e implementado para determinar el modificador que acompañará a la expresión numérica simplificada final que generará el sistema.

5.1.2.1. Obtención del candidato

A partir de la información anotada de la expresión numérica original, vamos a calcular el mejor candidato para ser usado en la versión simplificada de la expresión numérica. Usamos el valor dado (Vg), y a su normalización entre 0 y 1 la vamos a llamar *mapping given value* (Vmg), que representa la proporción que queremos simplificar. Recordemos que Vg no siempre está normalizado, sólo para los casos en los que la expresión numérica es un porcentaje. Para obtener la lista de candidatos usamos el *programa de aproximación de proporciones* (presentado en la sección 3.3.2), el cual a partir del valor de entrada (Vmg) devuelve una lista de candidatos para la sustitución.

La Figura 5.3 muestra un ejemplo de la salida del sistema calculada para el valor de entrada 0.28 correspondiente al porcentaje 28 %. La lista se organiza según los tipos de candidatos que ofrece, porcentajes o fracciones, en orden decreciente de precisión con respecto al valor de entrada, siendo la primera opción la más precisa para el tipo deseado. De ahí que dependiendo del nivel de dificultad elegido en el sistema, se elija la primera opción como candidato a sustituir por la expresión original. A ese valor elegido lo llamamos *candidate substitute value* (Vc). Esto significa que se elige al candidato más preciso para el nivel de dificultad seleccionado. El programa también devuelve valores como *none* or *all* si el valor de entrada es cercano a 0 o a 1 respectivamente.

A partir del valor Vc, calculamos el valor normalizado del candidato que lo llamamos *rounded value* (Vr). Con estos datos en la siguiente fase, se aplican las reglas de simplificación correspondientes.

Podemos ver un ejemplo del proceso completo en la Figura 5.4 para el nivel de dificultad *Fraction* y la expresión numérica original *more than 28 %*. Con Vmg=0.28 el sistema elige como candidato Vc=3/10 con un valor normalizado Vr=0.3.

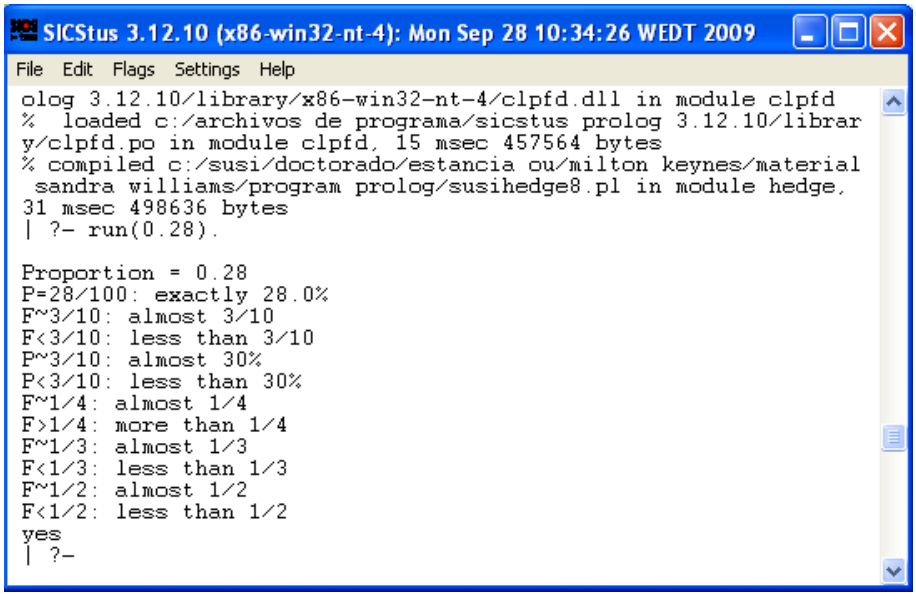


Figura 5.3: Salida del programa de aproximación de proporciones

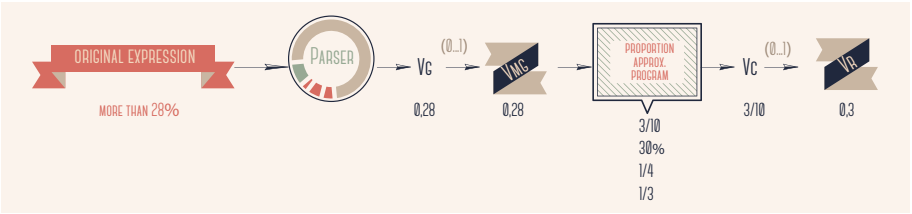


Figura 5.4: Proceso para obtener la expresión candidata para la simplificación. La expresión original *more than 28 %* es anotada por el *parser* (Vg), y este valor es normalizado (Vmg). Un valor candidato es elegido de la salida del *programa de aproximación de proporciones* (Vc) y es normalizado (Vr).

5.1.2.2. Aplicación de las reglas de simplificación

Recordamos las estrategias de simplificación de expresiones numéricas en inglés (presentadas en la sección 10.4.3) identificadas en nuestro estudio, a partir de las cuales se definen y se implementan las reglas del sistema computacional que presentamos.

Siguiendo la primera estrategia (las expresiones numéricas representadas en letras son transformadas por su correspondiente representación usando dígitos), el sistema aplicará una regla que se encargue de hacer esta transformación.

El uso del *programa de aproximación de proporciones* nos asegura que las estrategias seleccionadas en cada caso, porcentajes, fracciones o expresiones no numéricas, siguen las ideas que identificamos en el estudio empírico, que

decían que:

- Independientemente del nivel de dificultad para el que se está simplificando, las estrategias comunes identificadas son:
  - Los porcentajes se redondean al valor más próximo, tanto para valores comunes y no comunes, como para valores centrales y periféricos.
  - Las expresiones no numéricas se usan sólo para los valores extremos de la cantidad.
  - Las fracciones más comunes se usan en el rango central, mientras que en los rangos extremos se usan otro tipo de estrategias.

Además, tal como vimos en las estrategias identificadas que no había un comportamiento claro a la hora de utilizar los ratios como estrategia de simplificación, en nuestro sistema no se contempla esa posibilidad y no se usan ratios como candidatos de simplificación.

En el sistema se tiene en cuenta el nivel de dificultad que se ha elegido en la aplicación y de ahí se aplican las reglas correspondientes que siguen las ideas identificadas en el estudio:

- Teniendo en cuenta el nivel de dificultad para el que se está simplificando:
  - Si el nivel de dificultad se corresponde con el de una una persona que no entiende porcentajes, la estrategia a usar es cambiar las expresiones en porcentajes por sus correspondientes fracciones equivalentes.
  - Si el nivel de dificultad se corresponde con el de una persona que no entiende los porcentajes con decimales, la estrategia a usar es redondear la cantidad al porcentaje más próximo sin decimales.
  - Si el nivel de dificultad se corresponde con una persona que tiene dificultades con las expresiones numéricas de manera más general, las estrategias más usadas son fracciones. Adaptando las expresiones originales a este tipo de representación matemática, se simplifica su dificultad.

Para completar el proceso de simplificación entra el juego la decisión de cuándo usar o no modificador en la expresión simplificada. El sistema aplica un conjunto de reglas para decidir en cada caso, que veremos más adelante. Recordamos las conclusiones obtenidas de nuestro estudio empírico:

1. Independientemente del nivel de dificultad para el que se está simplificando, el uso de modificador no está influenciado ni por el valor central o periférico de la proporción, ni por la estrategia de simplificación utilizada.

2. Si no hay pérdida de precisión en la simplificación, entonces no se usa modificador.

3. No hemos encontrado correlación entre la pérdida de precisión y el uso de modificadores.

4. Hemos observado que si la expresión original tiene modificador, entonces en la expresión simplificada se mantiene.

Como no encontramos correlación entre la pérdida de precisión y el uso de modificadores, se decidió que el sistema iba a usar modificadores si había pérdida de precisión. Podremos observar esta decisión en el conjunto de reglas que se definieron para el uso de modificadores en la versión simplificada.

A partir de este conjunto de estrategias identificadas, las reglas que implementamos en nuestro sistema llevan a cabo la simplificación de expresiones numéricas en el texto de entrada dependiendo del nivel de dificultad elegido en la aplicación. Para cada expresión identificada, el sistema sólo aplica las reglas de simplificación si el nivel de dificultad de la expresión es mayor que el nivel de dificultad elegido. Siguiendo esta idea y teniendo en cuenta las conclusiones obtenidas en el estudio empírico del capítulo 4, definimos dos reglas de simplificación centradas en el tipo de expresión usada:

- Si el tipo de expresión numérica es *cardinal* o *fraction* y el formato es *words*, entonces el candidato para ser usado en la versión simplificada es el mismo número. Es decir, la expresión numérica se expresará en dígitos en lugar de en letras. Por ejemplo, si la expresión numérica original es *six*, será reemplazado por *6*, o si es *a quarter* será reemplazado por *1/4*.

- Si el tipo de expresión numérica es *percentages* o *decimal percentages* y el formato es *digit*, el candidato es calculado por el *programa de aproximación de proporciones* dado que el nivel de dificultad elegido en la interfaz del sistema es siempre menor que el nivel de dificultad de la expresión numérica.

Para completar el proceso de simplificación el sistema tiene que decidir si un modificador debería ser usado en la versión final de la expresión numérica simplificada que calcula. Esta decisión es tomada basada en la diferencia entre el valor de la expresión numérica original en el texto (Vg) y el valor del candidato a sustitución (Vc). En realidad se trabaja con los valores normalizados de dichos valores, Vmg y Vr, calculados en la primera fase de este proceso. También se tiene en cuenta si existe modificador en la expresión original.

Las posibles combinaciones de estos tres factores y su correspondiente elección del modificador para la versión simplificada se recogen en la Tabla

Expresión original	si Vr>Vmg	si Vr=Vmg	si Vr<Vmg
more than *expr*	around Vc	more than Vc	more than Vc
exactly *expr*	less than Vc	exactly Vc	more than Vc
less than *expr*	less than Vc	less than Vc	around Vc
expr	around Vc	Vc	around Vc

Tabla 5.1: Reglas para seleccionar el modificador. Para cada expresión original, los valores normalizados *(Vmg, Vr)* son usados para determinar el modificador elegido para la expresión simplificada. La versión final está compuesta por el modificador elegido y el valor del candidato seleccionado *(Vc)*

5.1. Aquí podemos ver todas las posibles opciones para decidir en cada caso el modificador y el valor correspondiente a la versión final. Por ejemplo, si la expresión original es *more than 28 %* y el nivel de dificultad es *Fraction*, tras la primera fase del proceso tenemos los siguientes valores Vg=0.28, Vmg=0.28, Vc=3/10 y Vr=0.3. Estamos entonces en el caso donde Vr>Vmg, primera columna de la Tabla 5.1, y como la expresión numérica original tenía modificador, estamos en la primera fila. Por tanto el modificador elegido para la versión simplificada es *around* acompañado de Vc, por lo que en nuestro caso, la expresión numérica simplificada sería *around 3/10*.

Por ejemplo, cuando en la interfaz del sistema el usuario elige el nivel de dificultad *Fractions*, cada porcentaje será reemplazado por una expresión numérica en forma de fracción calculada por el *programa de aproximación de proporciones*. Para cada expresión numérica original, dependiendo de los valores *Vr* y *Vmg* que calcula el sistema, se elige el modificador que acompañará a la versión simplificada a partir de las reglas definidas.

5.1.3. Etapa 4: Regeneración del texto

Esta etapa final de composición del texto es la misma que en el modelo genérico, es decir, a partir de las unidades lingüísticas simplificadas más el resto del texto original se genera una versión simplificada del texto original. Por lo tanto, la salida de nuestro sistema es una versión del texto original donde cada una de las expresiones numéricas originales han sido reemplazadas por su correspondiente versión simplificada.

A partir del texto original que presentamos en la sección 5.1 y que a continuación recordamos, los siguientes textos son los que obtenemos al aplicar cada etapa de nuestro modelo para simplificar las expresiones numéricas presentes en el texto. El primero corresponde al nivel de *Percentages without decimals (PWD)* y el segundo al nivel de *Fractions*. Cada expresión numérica simplificada calculada por el sistema es sustituida en el texto original.

Texto original

Another record year for A-levels

The A-level pass rate rose for the 26th year in a row as record number of teenagers achieved top grades. But figures released by the exam boards highlighted startling discrepancies in Grade A pass rates between regions across England. Statistics from the exam boards showed greater improvements in students in the South East getting A grades in the past **six** years than those in the North East. The South East has seen a **6.1 %** increase in A grades - to **29.1 %** - since 2002 but the North East has seen an improvement of only **2.1 %** - to **19.8 %** - during the same period. But the percentage of pupils gaining passing E grades is rising quicker in the North East - an improvement of **3.4 %** in **six** years compared with **2.8 %** in the South East. Overall figures showed the national pass rate soared **above 97 %** for the first time this year, while **one in four** **sixth**-formers were awarded A grades (**25.9 %**, up from **25.3 %** last year). The figures showed traditional subjects are still firm favourites with English and maths the top choices for candidates. Dr Mike Cresswell, director general of the AQA, said A-levels remained a "highly-valued qualification". He said he was particularly pleased to see the numbers of maths candidates rise from **60,093** last year to **64,593** this year. "There was an upward trend that began a couple of years ago that has accelerated. There are more candidates doing mathematics than at any time in the past. It's important we have people with high mathematic skills so that has to be good news."

Texto simplificado a nivel de *PWD*

Another record year for A-levels

Last Updated: Thursday, 14 August 2008, 08:28 GMT The A-level pass rate rose for the 26th year in a row as record number of teenagers achieved top grades. But figures released by the exam boards highlighted startling discrepancies in Grade A pass rates between regions across England. Statistics from the exam boards showed greater improvements in students in the South East getting A grades in the past **6** years than those in the North East. The South East has seen a **around 6 %** increase in A grades - to **around 29 %** - since 2002 but the North East has seen an improvement of only **around 2 %** - to **around 20 %** - during the same period. But the percentage of pupils gaining passing E grades is rising quicker in the North East - an improvement of **around 3 %** in **6** years compared with **around 3 %** in the South East. Overall figures showed the national pass rate soared **above 97 %** for the first time this year, while **1/4** **6th**-formers were awarded A grades (**around 26 %**, up from **around 25 %** last year) The figures showed traditional subjects are still firm favourites with English and maths the top choices for candidates. Dr Mike Cresswell,

director general of the AQA, said A-levels remained a "highly-valued qualification". He said he was particularly pleased to see the numbers of maths candidates rise from **60,093** last year to **64,593** this year. "There was an upward trend that began a couple of years ago that has accelerated. There are more candidates doing mathematics than at any time in the past. It's important we have people with high mathematic skills so that has to be good news."

Texto simplificado a nivel de *Fractions*
Another record year for A-levels
Last Updated: Thursday, 14 August 2008, 08:28 GMT The A-level pass rate rose for the 26th year in a row as record number of teenagers achieved top grades. But figures released by the exam boards highlighted startling discrepancies in Grade A pass rates between regions across England. Statistics from the exam boards showed greater improvements in students in the South East getting A grades in the past **6** years than those in the North East. The South East has seen a **around 1/10** increase in A grades - to **around 3/10** - since 2002 but the North East has seen an improvement of only **around none** - to **around 1/5** - during the same period. But the percentage of pupils gaining passing E grades is rising quicker in the North East - an improvement of **around none** in **6** years compared with **around none** in the South East. Overall figures showed the national pass rate soared **around all** for the first time this year, while **1/4** **6**th-formers were awarded A grades (**around 1/4**, up from **around 1/4** last year) The figures showed traditional subjects are still firm favourites with English and maths the top choices for candidates. Dr Mike Cresswell, director general of the AQA, said A-levels remained a "highly-valued qualification". He said he was particularly pleased to see the numbers of maths candidates rise from **60,093** last year to **64,593** this year. "There was an upward trend that began a couple of years ago that has accelerated. There are more candidates doing mathematics than at any time in the past. It's important we have people with high mathematic skills so that has to be good news."Last Updated: Thursday, 14 August 2008, 11:28 GMT

5.2. Evaluación del sistema de simplificación de expresiones numéricas en inglés

Para saber como de bueno es nuestro sistema y las mejoras que se pueden hacer, realizamos una evaluación del mismo. En esta sección describimos el material, el experimento, los participantes que formaron parte de la evaluación y los resultados obtenidos.

5.2.1. Materiales para la evaluación del sistema

Hemos seleccionado para el experimento un conjunto de ocho oraciones candidatas del corpus del proyecto *NumGen*. El número de expresiones numéricas involucradas es mayor, ya que algunas oraciones contienen más de una expresión. En total tenemos 13 expresiones numéricas. Las oraciones fueron seleccionadas con la mayor variación posible de contexto y de precisión.

El rango de los valores de las proporciones va desde el 0,0 hasta el 1,0 para dar la mayor cobertura posible a los valores de las expresiones candidatas. Consideramos valores en el rango central (aquellos que estén entre el 0,2 y el 0,8) y valores que están en los extremos (del 0,0 al 0,2 y del 0,8 al 1,0).

Además, clasificamos los valores como comunes y no comunes según su frecuencia de uso. Por ejemplo, fracciones como *1/4* o *1/2* y porcentajes como *25 %* o *50 %* son comúnmente más usados en comparación con otros como *4/7* o *13 %*.

5.2.2. Experimento para evaluar el sistema

Para evaluar el sistema, presentamos un cuestionario a un conjunto de evaluadores humanos. El experimento fue diseñado e implementado en SurveyMonkey[1], un proveedor de uso general para encuestas online.

Para cada oración original presentamos dos posibles oraciones generadas por el sistema. A los participantes se les preguntó que, según su propio juicio, decidieran si estaban de acuerdo con las oraciones simplificadas para la original dada. Usamos una escala Likert de cuatro valores (*Strongly Disagree*, *Disagree*, *Agree*, *Strongly Agree*) para recoger las respuestas.

En el cuestionario se presentaron sólo dos niveles de adaptación de la oración original. La primera opción era generada por el sistema cuando se seleccionaba el nivel *Fractions* y la segunda opción cuando el nivel era *Percentages without decimals (PWD)*.

5.2.3. Participantes del experimento

Les pedimos a los mismos expertos con los que habíamos realizado la identificación experimental que evaluaran nuestro sistema (sección 4.4.3). Estamos muy agradecidos a todos los participantes por su implicación en el experimento, primero para identificar las estrategias y después para evaluar la simplificación automática llevada a cabo por el sistema que hemos implementado.

[1] http://www.surveymonkey.com/s/WJ69L86

Nivel	Media Total	Valores	Media	Valores	Media
Fractions	2,44	Centrales	2,87	Comunes	2,59
		Extremos	2,14	No comunes	1,21
PWD	2,96	Centrales	3,00	Comunes	2,80
		Extremos	2,96	No comunes	3,22

Tabla 5.2: Evaluación del sistema: nivel de *Fractions* y nivel de *Percentages without decimals (PWD)*

5.2.4. Resultados de la evaluación del sistema

Las respuestas de los participantes fueron analizadas y evaluadas. En total recogimos 377 respuestas, 191 de ellas para la primera opción presentada (nivel *Fractions*) y 186 respuestas para la segunda opción presentada (nivel *Percentages without decimals (PWD)*). La Tabla 5.2 muestra la valoración para los valores centrales y extremos, y para los comunes y no comunes, para cada una de las opciones presentadas en la encuesta. Los valores se presentan en medias aritméticas de las respuestas recogidas, considerando el valor 1 con la opción de totalmente en desacuerdo (*Strongly Disagree*) y el valor 4 con totalmente de acuerdo (*Strongly Agree*), según la escala presentada en el cuestionario.

Para el nivel de *Fractions*, hay diferencia entre la media de los valores centrales y la media de los valores extremos y presenta una media claramente más alta con respecto a los valores comunes y no comunes. En el nivel *Percentages without decimals (PWD)* hay más diferencia entre las medias de comunes y no comunes que entre los valores centrales y extremos.

A partir de los datos recogidos hemos observado que para las fracciones los participantes están más de acuerdo con las simplificaciones realizadas por el sistema para los valores comunes y centrales. En cambio, para los porcentajes sin decimales no hay distinción entre los distintos casos, ya que lo que prima es eliminar los decimales aproximando la cantidad a la más cercana, independientemente de que sea central o extrema, común o no común.

Además, en la Tabla 5.3 mostramos la distribución en porcentajes de la opinión de los expertos para cada una de las opciones presentadas. En la primera opción, nivel *Fractions*, no hay demasiada diferencia entre la opinion de los expertos que están de acuerdo con la salida del sistema y con los que están en desacuerdo: un 54 % (43 % + 11 %) está de acuerdo, frente a un 46 % (19 % + 27 %) que está en desacuerdo. Pero para la segunda opción presentada, nivel *Percentages without decimals (PWD)*, la opinión de los expertos muestra que la mayoría están de acuerdo con las simplificaciones hechas, ya que un 79 % (56 % + 23 %) está de acuerdo frente al 21 % (6 % + 15 %). Podemos observar la opinión de los expertos en los gráficos de

Opinión	Nivel *Fractions*	Nivel *PWD*
Totalmente en desacuerdo	19 %	6 %
Desacuerdo	27 %	15 %
De acuerdo	43 %	56 %
Totalmente de acuerdo	11 %	23 %

Tabla 5.3: Porcentaje de los expertos para cada opción en ambos niveles del sistema

porcentajes que se presentan en la Figura 5.5 y en la Figura 5.6.

En términos generales, los expertos piensan que en el nivel *PWD* las simplificaciones hechas por el sistema son mejores que las hechas en el nivel *Fractions*. Los expertos están además especialmente en desacuerdo cuando la simplificación usa fracciones en dos casos: uno es en el tratamiento de valores extremos donde el sistema obtiene como posibles candidatos *none* y *all*. Por ejemplo, para la expresión *1.3 %* el sistema la simplificaba usando *around none*. Otro caso es cuando fracciones no comunes son usadas para simplificar expresiones numéricas, como por ejemplo, cuando la expresión *87.8 per cent* se simplifica usando *around 9/10*. En estos dos casos podemos observar que la media es más baja que en el resto de los valores recogidos.

5.2.5. Discusión de los resultados

El sistema presentado combina transformaciones sintácticas (introduciendo modificadores) y sustituciones léxicas (reemplazando valores actuales por alternativas más adecuadas y transformando cantidades expresadas en palabras por dígitos) para simplificar la expresión numérica original. Las reglas definidas en nuestro sistema son de uso específico y centradas en la simplificación de expresiones numéricas. Con este tipo de transformaciones pretendemos mejorar la legibilidad del texto a pesar del hecho de que la estructura sintáctica resultante de la expresión numérica simplificada sea más complicada debido a la presencia de modificadores. En la Figura 5.7 podemos ver los árboles sintácticos correspondientes al ejemplo de la expresión numérica original *25.9 %* y la expresión simplificada que genera el sistema *more than a quarter*, donde podemos ver que la versión simplificada es sintácticamente más compleja. Esto puede influir a la hora de leer y comprender un texto para determinados usuarios reales, como pudimos observar para el caso de las personas con dislexia en textos en español, por lo que como trabajo futuro habría que comprobarlo para usuarios reales para el caso del inglés.

Con respecto a la cobertura de los diferentes tipos de expresiones numéricas, el sistema no considera los *ratios* como una posible estrategia de

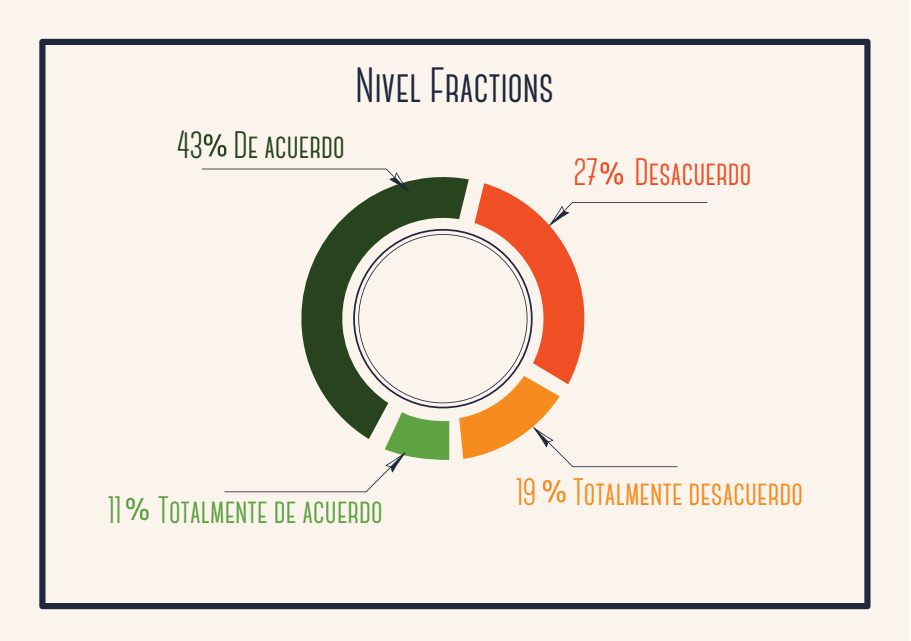


Figura 5.5: Gráfico de porcentajes de la opinión de los expertos en el nivel de fracciones en el sistema para el inglés

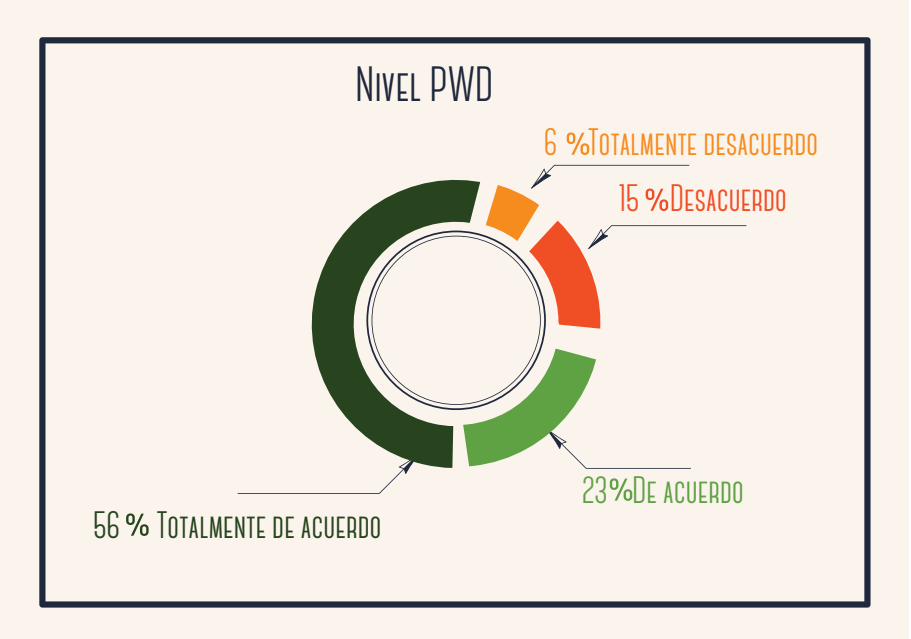


Figura 5.6: Gráfico de porcentajes que recoge la opinión de los expertos en el nivel de porcentajes sin decimales en el sistema para el inglés

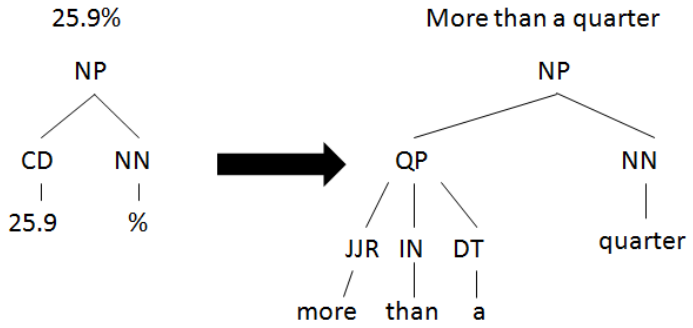


Figura 5.7: Árboles sintácticos correspondientes a la expresión numérica original y su correspondiente simplificación

simplificación, ya que el *programa de aproximación de proporciones* que usamos no usa los ratios como candidatos para simplificar una proporción de entrada. Esta posibilidad podría ser explorada en un futuro trabajo de ampliación del sistema.

Otra observación es que el sistema no tiene en cuenta el contexto de la expresión numérica que aparece en la frase. Por ejemplo, si la oración hace una comparación entre dos expresiones numéricas que el sistema redondea con el mismo valor, el significado original se ha perdido. Un ejemplo de este caso lo podemos ver en el texto de ejemplo que hemos usado, que contiene la siguiente frase "*But the percentage of pupils gaining passing E grades is rising quicker in the North East - an improvement of **3.4 %** in six years compared with **2.8 %** in the South East.*". Ambos porcentajes *3.4 %* y *2.8 %* son simplificados por el sistema usando *around 3 %* y el significado de la oración original se pierde, ya que la comparativa entre las dos expresiones no se mantiene en la versión simplificada "*But the percentage of pupils gaining passing E grades is rising quicker in the North East - an improvement of **around 3 %** in 6 years compared with **around 3 %** in the South East.*". Estos casos nos hacen plantearnos la importancia del contexto a la hora de llevar a cabo la simplificación de expresiones numéricas para que se conserve el significado original.

5.3. Sistema de simplificación de expresiones numéricas en español

El objetivo de este segundo sistema es simplificar expresiones numéricas en textos en español. Siguiendo de nuevo el modelo genérico presentado en la sección 4.1, diseñamos e implementamos un sistema que aplica las reglas aprendidas de la identificación experimental realizada para generar un tex-

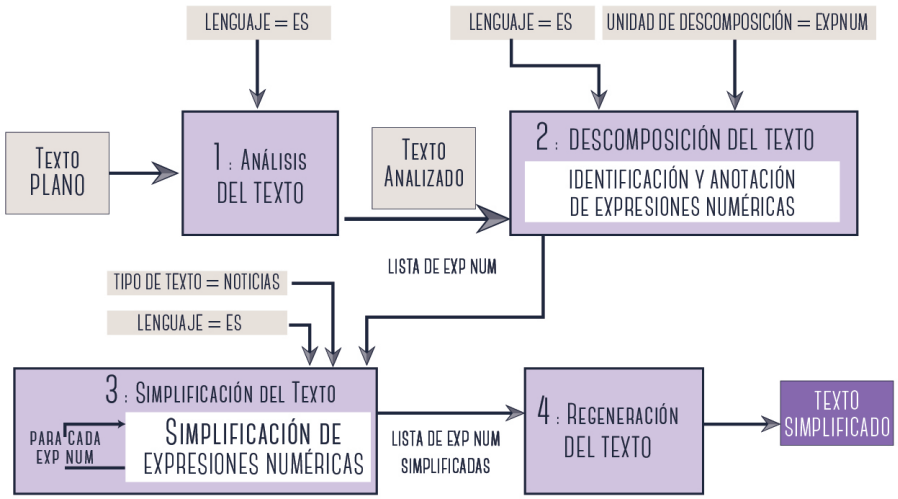


Figura 5.8: Etapas del modelo automático de simplificación centrado en expresiones numéricas para el español tal y como se ha instanciado para el sistema de simplificación de expresiones numéricas en español

to en español con las expresiones numéricas simplificadas. Nuestro sistema está compuesto por distintos componentes que se encargan de cada etapa y son integrados en un plug-in desarrollado en Java y que se usa desde la herramienta GATE. El sistema fue presentado en los trabajos de Bautista y Saggion (2014a) y Bautista y Saggion (2014b).

En las siguientes secciones explicamos cómo se trabaja en cada etapa y las decisiones que se tienen que tomar para realizar el proceso de simplificación. En la Figura 5.8 podemos ver cómo se instancia para este sistema el modelo de simplificación de expresiones numéricas (presentado en la Figura 4.2).

Mostramos un ejemplo de un texto del corpus del proyecto *Simplext* (Saggion et al., 2011), obtenido de la Agencia de Noticias Servimedia[2]. Podemos ver que el texto consiste en 5 oraciones y contiene 10 tipos diferentes de expresiones numéricas, incluyendo porcentajes, números enteros, decimales, expresiones con y sin modificadores, etc.

> **CASI EL 20 %** DE LAS AGRESIONES QUE SUFREN LOS MÉDICOS CAUSAN LESIONES
>
> El **18,55 %** de las agresiones que sufrieron los médicos españoles en sus consultas el año pasado tuvieron como consecuencia una lesión, seuán los datos de el Observatorio de Agresiones de la Organización Médica Colegial, que indican también que el **13,4 %** de los facultativos afectados por esta situación pidieron por esta causa la baja laboral.

[2]http://www.servimedia.es

> En virtud de estas cifras, difundidas este martes en rueda de prensa, en 2010 se registraron en España un total de **451** agresiones a facultativos, es decir, **2,07** por cada **mil** médicos, lo que supone, a juicio de la organización médica, un "grave problema social" para el que se pide "tolerancia cero" y que se produce en el **90,63 %** de los casos en el sector público.
> El ámbito médico más afectado por las agresiones de pacientes, es, en virtud del observatorio creado por los colegios de facultativos, el de Atención Primaria, donde se contabilizaron en 2010 el **65 %** de los atentados a profesionales sanitarios.
> Y el grupo de edad más castigado, el que va desde los **46** a los **55** años.

5.3.1. Etapa 1: Análisis del texto

La primera variable que hay que tener en cuenta es el lenguaje del texto original que se quiere simplificar, y que en este caso es el español. En esta etapa de análisis se realizan dos tareas: por un lado, el etiquetado de las categorías gramaticales del texto (*part-of-speech tagging*) y por otro lado, el análisis sintáctico del texto elegido que se guarda en formato XML.

La herramienta elegida para ambas tareas es FreeLing (Padró et al., 2010). Es una de las herramientas de análisis para el castellano que, entre otros, permite realizar el etiquetado de las categorías gramaticales basado en un modelo de Markov con estados ocultos. Este tipo de etiquetado 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[3]. Para el propósito de este trabajo 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 singular (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 expresiones correspondientes a 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).

el	el	DA0MS0	1
65 %	65/100	Zp	1
de	de	SPS00	0.999919
los	el	DA0MP0	0.97623
atentados	atentado	NCMP000	0.75
va	ir	VMIP3S0	1
desde	desde	SPS00	1
los	el	DA0MP0	0.97623
46	46	Z	1
a	a	SPS00	0.99585
los	el	DA0MP0	0.97623
55	55	Z	1
años	año	NCMP000	1

Tabla 5.4: Ejemplo del análisis morfológico obtenido por FreeLing

Para cada texto se generan dos archivos. El primero de ellos, con extensión *.tagged*, guarda el etiquetado léxico que asigna a cada palabra: su lema, su categoría morfológica y la probabilidad de que la etiqueta asociada a ese token sea correcta. El segundo archivo, con extensión *.xml*, almacena el análisis sintáctico realizado a nivel de oración para el texto de entrada. Para cada palabra se le asignan una serie de atributos como son un identificador, la etiqueta morfológica, el lema y la forma. De los dos archivos generados por FreeLing, nuestro sistema trabaja con el *.xml* para añadir la información extra que anota en una de las etapas del proceso.

Como ejemplo, para ver como FreeLing asigna las categorías morfológicas correspondientes, mostramos en la Tabla 5.4 una parte del archivo *.tagged* donde podemos ver un par de expresiones numéricas pertenecientes a un par de oraciones del texto presentado como ejemplo: "*...el 65 % de los atentados...*" y " *... va desde los 46 a los 55 años...*". Se pueden ver las etiquetas obtenidas, donde la primera columna es la palabra de la oración, la segunda columna es el lema de la palabra, la tercera es la etiqueta correspondiente a la categoría morfológica asignada por FreeLing y la cuarta columna es la probabilidad calculada por el analizador.

La información del análisis sintáctico realizado se guarda en un archivo *.xml*. A continuación, mostramos un pequeño ejemplo de parte de una oración *"...el 65 % de los atentados..."*, donde podemos ver que a cada palabra se le asigna un conjunto de atributos como son: un identificador (*id*), su posición (*span*), la etiqueta morfosintáctica (*tag*), el lema (*lemma*), la forma (*form*) y un par de atributos más que calcula FreeLing.

[3] http://nlp.lsi.upc.edu/freeling/doc/tagsets/tagset-es

```
<word id="179" span="993-995" guess="no" recov="el"
 tag="DA0MS0" lemma="el" form= "el">
    el
</word>
<word id="180" span="995-999" guess="yes"recov="65 %"
 tag="Zp" lemma="65/100" form= "65 %">
    65 %
</word>
<word id="181" span="1000-1002" guess="no" recov="de"
 tag="SPS00" lemma="de" form= "de">
    de
</word>
<word id="182" span="1003-1006" guess="no" recov="los"
 tag="DA0MP0" lemma="el" form= "los">
    los
</word>
<word id="183" span="1007-1016" guess="no" recov="atentados"
 tag="NCMP000" lemma="atentado" form="atentados">
    atentados
</word>
```

La salida de esta etapa es una lista de oraciones analizadas donde cada palabra ha sido etiquetada con su correspondiente categoría morfológica. Este análisis es utilizado en la siguiente etapa para llevar a cabo la identificación y anotación de las expresiones numéricas presentes en el texto.

5.3.2. Etapa 2: Descomposición del texto

En esta etapa se realiza la descomposición del texto en las unidades objetivo que correspondan. En un primer paso descomponemos el texto en oraciones, comenzando por la oración correspondiente al titular de la noticia y continuando con las oraciones del cuerpo del texto.

Posteriormente, para cada oración tenemos que encontrar y recolectar las unidades objetivo para el proceso de simplificación, en nuestro caso, las expresiones numéricas. En esta estapa están involucrados dos pasos diferentes. El primero de ellos será identificar las expresiones candidatas al proceso de simplificación a partir de los resultados del análisis léxico-sintáctico realizado en la etapa anterior. El segundo será la anotación de las expresiones con la información necesaria para el proceso de simplificación.

5.3.2.1. Etapa 2.1: Identificación de expresiones numéricas

En esta etapa identificamos diferentes tipos de expresiones numéricas en el texto original a partir del análisis realizado en las etapas previas. La información numérica puede estar expresada en representación matemática como es el caso de los porcentajes (*78 %*, *23 por ciento*), fracciones (*1/4*, *un tercio*), ratios (*18 de cada 100*, *1 de cada 8*), etc. Dependiendo del lenguaje del

Original	Forma	Lema	Etiqueta
854	854	854	Z
trescientos quince	trescientos_quince	315	Z
un millón	un_millón	1000000	Zd
una docena	una_docena	12	Zd
uno de cada cuatro	uno_de_cada_cuatro	1/4	Zp
tres octavas partes	tres_octavas_partes	3/8	Zp
ochenta por ciento	ochenta por ciento	80/100	Zp
36 %	36_%	36/100	Zp
300 dólares	300_dólares	$_USD:300	Zm
20 km por hora	20_km_por_hora	SP_km/h:20	Zu
8 gramos/litro	8_gramos_/_litro	DN_g/l:8	Zu

Tabla 5.5: Ejemplos de cómo analiza FreeLing los numerales

texto original y de las herramientas usadas en el proceso de análisis de las etapas anteriores, las expresiones numéricas del texto pueden ser identificadas de diferentes formas, a partir de las etiquetas morfológicas usadas por el analizador, usando gramáticas, a partir de un conjunto de reglas, etc.

En nuestro caso, el análisis llevado a cabo por FreeLing nos permite identificar diferentes tipos de expresiones numéricas etiquetadas siguiendo el sistema de clasificación del analizador para la información numérica. Ya vimos que el analizador utiliza un conjunto de etiquetas para representar la información morfológica de los numerales basadas en la etiquetas propuestas por el grupo EAGLES para la anotación morfosintáctica de lexicones y corpus para todas las lenguas europeas. En la Tabla 5.5 podemos ver algunos ejemplos del análisis que realiza FreeLing. Para cada expresión numérica, identifica: su forma (si tiene más de un token, los une con guiones bajos), su lema (aplicando el lematizador del analizador) y su etiqueta correspondiente.

5.3.2.2. Etapa 2.2: Anotación de expresiones numéricas

En esta etapa extraemos información de las expresiones numéricas identificadas en la etapa anterior con el objetivo de llevar a cabo la anotación de la información extra necesaria en el proceso de simplificación.

El proceso de extracción de información consiste en identificar los diferentes componentes de una expresión numérica:

1. Modificadores: *casi*, *más de* o *alrededor de*, entre otros.
2. Cantidad: *500*, *23 %* ó *6,7*.
3. Tipo de representación matemática: porcentajes, fracciones, números decimales, ratios, etc.

4. Unidades de medida: *km*, *litro* o *millones de euros*.

5. Tipo de expresión numérica, según la clasificación que estamos usando del analizador: partitiva, monetaria, etc.

En nuestro caso, para anotar las diferentes expresiones numéricas en los textos originales incluyendo sus distintos componentes, hemos utilizado GATE para definir un conjunto de gramáticas JAPE (*Java Annotation Patterns Engine*)(sección 3.3.3). 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 que 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.

La Tabla 5.6 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 (LHS), y la parte derecha (RHS) 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 ejemplo, 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.

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

Tabla 5.6: Ejemplo de una regla de una gramática JAPE

Etiqueta	Expresión Numérica	Ejemplo
CASIporcFract	casi + Zp	casi un cuarto
DURANTENUM	durante + Z	durante 24 días
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...

Tabla 5.7: Tipos identificados en el corpus usado para medir la cobertura de las reglas

Estas gramáticas JAPE las usamos para anotar los distintos tipos de expresiones numéricas. Para dar cobertura a los distintos tipos de expresiones numéricas que nos encontramos en los textos hemos definido 45 reglas. Para desarrollarlas 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 búsquedas 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.

Una vez implementadas las reglas, se llevó a cabo una corrección manual de las mismas para evaluar la cobertura que proporcionaban las reglas. Para ello se seleccionó un subconjunto de 10 textos con un total de 59 oraciones. Usando la herramienta GATE se hace una comparación automática identificando las etiquetas nuevas creadas manualmente y las generadas automáticamente a partir de las gramáticas JAPE definidas. Sólo seis etiquetas tuvieron que ser mejoradas de las definidas inicialmente, lo que se corresponde con un 10,52 % de error del conjunto de reglas definido. En el corpus seleccionado sólo aparecían expresiones numéricas de 13 de los 45 tipos representados por las reglas. En la Tabla 5.7 mostramos los 13 casos identificados en el corpus usado para medir la cobertura de las reglas definidas.

Además, 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 los tres valores y hemos observado que en las expresiones numéricas menos frecuentes se obtienen peores resultados, pero en general para las expresiones numéricas más frecuentes se obtienen muy buenos. En los resultados globales hemos visto que

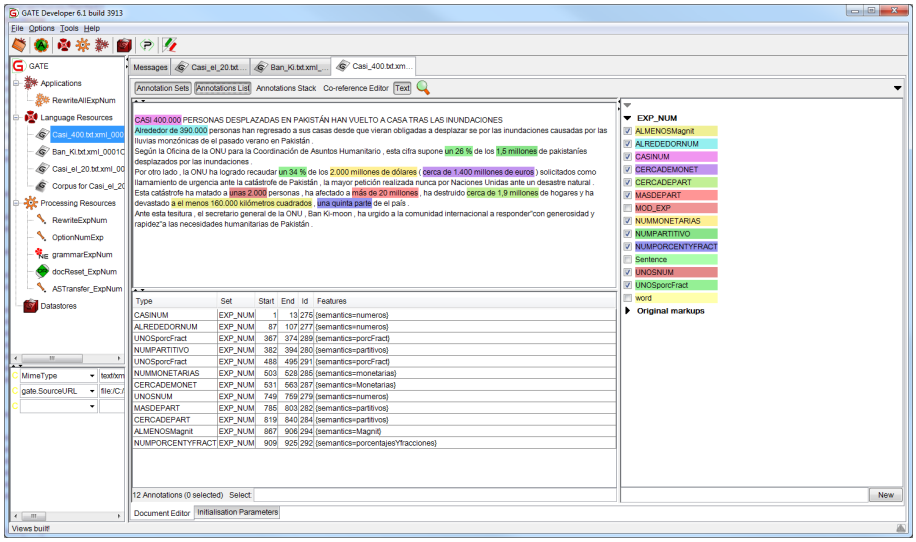


Figura 5.9: Anotación de expresiones numéricas en GATE

tenemos una *precision* y un *recall* muy altos, ya que nuestras reglas etiquetan una fracción bastante alta de las instancias relevantes del corpus.

La salida de esta etapa de anotación de expresiones numéricas es una lista de expresiones numéricas anotadas con toda la información necesaria para el proceso de simplificación que se realiza en la siguiente etapa. La Figura 5.9 muestra la interfaz de GATE donde podemos ver las anotaciones de las distintas expresiones numéricas del texto producidas a partir de aplicar las gramáticas JAPE definidas.

5.3.3. Etapa 3: Simplificación del texto

La etapa de simplificación del texto recibe la lista de expresiones numéricas identificadas y anotadas en la etapa anterior. A partir de ahí tenemos que realizar la simplificación de cada una de las expresiones numéricas que aparecen en el texto original para generar una versión simplificada como salida de nuestro sistema. Para llevar a cabo esta simplificación se definen e implementan un conjunto de reglas a partir de las estrategias de simplificación identificadas para el español.

Nuestro sistema considera de manera general las siguientes estrategias que se implementan en un conjunto de reglas computacionales:

1. Las expresiones representadas en letras se cambian por expresiones representadas en dígitos.

2. Si la expresión numérica original tiene modificador, éste se deja en la versión simplificada y la cantidad se redondea.

3. En cambio, si la expresión numérica original no tiene modificador, aplicando una serie de reglas, un modificador se elige y se añade junto con la cantidad redondeada.

Hay que señalar que del conjunto de estrategias identificadas en el estudio empírico llevado a cabo, un subconjunto de ellas no es implementado por el sistema y no es considerado en el proceso automático de simplificación de expresiones numéricas en español en nuestro caso. Las estrategias a las que nos referimos son:

- La eliminación de expresiones numéricas. Preferimos simplificar la información que perderla en el proceso de generación de la versión simplificada.
- La reescritura de las expresiones numéricas usando información del contexto. Actualmente nuestro sistema no dispone de ningún mecanismo automático para llevar a cabo la reescritura de la expresión original por una versión simplificada de la misma que use información del contexto.

Las reglas de simplificación implementadas siguen el siguiente proceso: la cantidad es siempre redondeada y un conjunto de reglas es aplicado para elegir el modificador teniendo en cuenta la pérdida de precisión. Para obtener el número redondeado correspondiente a la cantidad original, se realizan cálculos matemáticos usando diferentes métodos del paquete *Math* de Java, que nos permite redondear la cantidad al entero más próximo por encima a partir de la cantidad original. Por ejemplo, si el valor original de la cantidad es *0,891*, el sistema calcula el valor redondeado *1,0*. Si en la expresión numérica original hay unidades, también son tratadas en el proceso de simplificación. La versión simplificada está compuesta por el modificador elegido, la cantidad redondeada y las unidades, si las hay.

Para elegir el modificador para la expresión simplificada definimos cuatro reglas. Si en la expresión original ya había modificador, se mantiene y la cantidad es redondeada. Para el resto de los casos, el sistema compara la cantidad original con la cantidad redondeada y dependiendo de los valores selecciona un modificador u otro. En la Tabla 5.8 podemos ver las reglas de selección de modificadores acompañadas de un ejemplo para cada caso.

Así, en esta etapa cada expresión numérica es simplificada usando el conjunto de reglas definidas e implementadas para el proceso de simplificación del sistema. Como resultado obtenemos una lista de expresiones numéricas simplificadas que usaremos en la siguiente etapa.

Expresión original	Cantidad redondeada	Caso	Modificador	Expresión simplificada
alrededor de 5689 millones	*6000 millones*	Hay un modificador en la expresión original	Modificador original se mantiene	*alrededor de 6000 millones*
27.3 %	*25 %*	No hay modificador en la expresión original, y original > redondeada	Se añade modificador 'más de'	*más de 25 %*
476	*500*	No hay modificador en la expresión original, y original < redondeada	Se añade modificador 'casi'	*casi 500*
3000	*3000*	No hay modificador en la expresión original, y original = redondeada	Sin modificador	*3000*

Tabla 5.8: Selección del modificador para la expresión numérica simplificada en diferentes casos. Cada caso viene acompañado de un ejemplo.

5.3.4. Etapa 4: Regeneración del texto

La etapa final de composición del texto es la misma que en el modelo genérico, es decir, a partir de las expresiones numéricas simplificadas, junto con el resto del texto, se genera una versión completa simplificada del texto original. La salida de nuestro sistema es por lo tanto una versión del texto original donde las expresiones numéricas originales han sido simplificadas a partir de unas reglas de simplificación aplicadas.

Además de reemplazar las expresiones numéricas se realiza un procesamiento posterior del texto para resolver algunos errores que se producen en el tratamiento del texto por el analizador FreeLing. Entre los errores está el caso de transformación de contracciones del español: *del* es separado en el análisis en dos componentes *de + el*, y *al* es separado en *a + el*. Estos casos deben ser reconstruidos para generar la versión final del texto. También hay que tratar las comillas, paréntesis, barras, guiones y otras marcas de puntuación, que el analizador separa introduciendo blancos.

A continuación podemos ver el texto original que usamos como ejemplo (sección 11.3), y el texto obtenido con las expresiones numéricas simplificadas al aplicar cada etapa de nuestro modelo en el sistema implementado para simplificar expresiones numéricas en español. Podemos ver como las cantidades originales se redondean y se utilizan diferentes modificadores dependiendo de la expresión numérica original. Por ejemplo, *18,55 %* se simplifica por la expresión *casi 19 %*, mientras que *13,4 %* se simplifica por la expresión *más de 13 %*, y así para el resto de los casos.

Texto original

CASI EL 20 % DE LAS AGRESIONES QUE SUFREN LOS MÉDICOS CAUSAN LESIONES

El **18,55 %** de las agresiones que sufrieron los médicos españoles en sus consultas el año pasado tuvieron como consecuencia una lesión, seuán los datos de el Observatorio de Agresiones de la Organización Médica Colegial, que indican también que el **13,4 %** de los facultativos afectados por esta situación pidieron por esta causa la baja laboral.

En virtud de estas cifras, difundidas este martes en rueda de prensa, en 2010 se registraron en España un total de **451** agresiones a facultativos, es decir, **2,07** por cada **mil** médicos, lo que supone, a juicio de la organización médica, un "grave problema social" para el que se pide "tolerancia cero" y que se produce en el **90,63 %** de los casos en el sector público.

El ámbito médico más afectado por las agresiones de pacientes, es, en virtud del observatorio creado por los colegios de facultativos, el de Atención Primaria, donde se contabilizaron en 2010 el **65 %** de los atentados a profesionales sanitarios.

Y el grupo de edad más castigado, el que va desde los **46** a los **55** años.

Texto simplificado
CASI EL 20 % DE LAS AGRESIONES QUE SUFREN LOS MÉDICOS CAUSAN LESIONES
El **casi 19 %** de las agresiones que sufrieron los médicos españoles en sus consultas el año pasado tuvieron como consecuencia una lesión, según los datos del Observatorio de Agresiones de la Organización Médica Colegial, que indican también que el **más de 13 %** de los facultativos afectados por esta situación pidieron por esta causa la baja laboral.
En virtud de estas cifras, difundidas este martes en rueda de prensa, en 2010 se registraron en España un total de **casi 500** agresiones a facultativos, es decir,**más de 2** por cada **1000** médicos, lo que supone, a juicio de la organización médica, un "grave problema social" para el que se pide "tolerancia cero" y que se produce en el **casi 91 %** de los casos en el sector público.
El ámbito médico más afectado por las agresiones de pacientes, es, en virtud del observatorio creado por los colegios de facultativos, el de Atención Primaria, donde se contabilizaron en 2010 el **más de 60 %** de los atentados a profesionales sanitarios.
Y el grupo de edad más castigado, el que va desde los **casi 50** a los **casi 60** años.

5.4. Evaluación del sistema de simplificación de expresiones numéricas en español

El sistema de simplificación automático de expresiones numéricas en español ha sido evaluado de dos maneras distintas. La primera de ellas fue una evaluación de manera automática para analizar la precisión lingüística de la salida del sistema. La segunda evaluación se realizó con expertos para que evaluaran directamente la salida del sistema. A continuación presentamos los detalles de cada una de las evaluaciones realizadas.

5.4.1. Evaluación automática

Para llevar a cabo esta evaluación hemos usado un subconjunto de textos del corpus perteneciente al proyecto *Simplext* (sección 3.1). El corpus usado para la evaluación tiene 57 textos, de los cuales 29 textos tienen expresiones numéricas en un total de 73 oraciones.

La finalidad era analizar la precisión lingüística de la salida del sistema, comprobando que la oración simplificada era correcta y que se preservaba el significado en el proceso de simplificación. Para ello se compararon la oración original y la simplificada según el criterio del evaluador, que en este caso fue la propia autora de esta tesis.

Los resultados que obtuvimos fueron que de las 73 oraciones, en 61 de

ellas se realizaba un reemplazo de la expresión numérica de manera que la oración seguía siendo correcta y se preservaba el significado de la oración original, frente a 12 oraciones donde el reemplazo no fue efectivo. De aquí obtuvimos que el 83,56 % (casi 84 %) de las oraciones simplificadas eran correctas y preservaban el significado.

Aunque este resultado es muy positivo, el análisis cualitativo de los resultados revela que existen algunos errores debidos a un mal postprocesamiento de la oración de salida o a un mal tratamiento de las expresiones numéricas que aparecen en una expresión comparativa. En el siguiente ejemplo podemos ver como dos cantidades numéricas son comparadas (*22.435 frente a 21.875*) y la versión simplificada para estas expresiones es la misma para ambas (*más de 20.000*), por lo que no se preserva el significado de la oración original.

La oración original es: *"Las cifras de disoluciones se mantienen en 2010 similares a las de 2009, 22.435 frente a 21.875, con un ligero incremento del 2,56 %"*, y la salida del sistema es *"Las cifras de disoluciones se mantienen en 2010 similares a las de 2009, más de 20000 frente a más de 20000, con un ligero incremento del casi 3 %"*.

En casos así se observa la importancia de tratar el contexto de las expresiones numéricas en la frase para determinar que ambas expresiones numéricas están relacionadas y considerarlo a la hora de simplificar las expresiones.

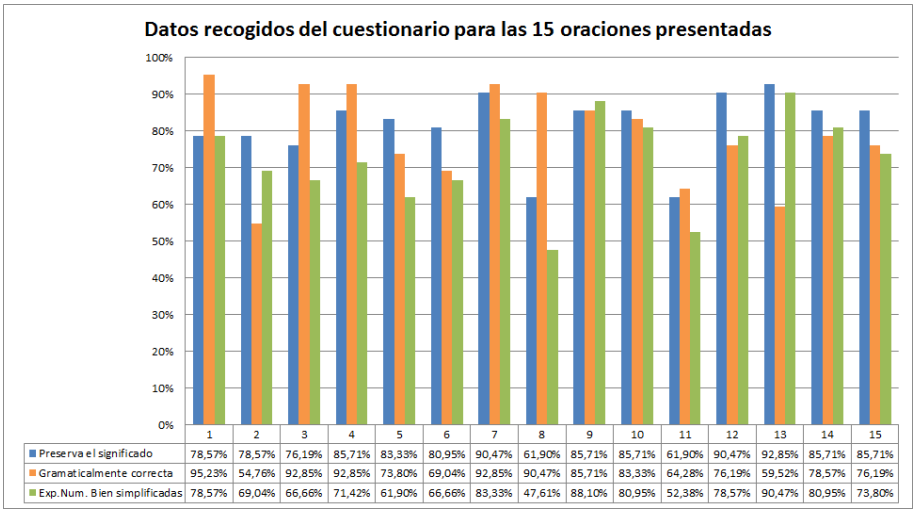
5.4.2. Evaluación con expertos

Para realizar esta evaluación contamos con la participación de expertos, profesores de primaria y secundaria que trabajan diariamente con alumnos que necesitan adaptaciones y que tienen formación académica para evaluar la simplificación hecha por nuestro sistema. En nuestra evaluación han participado 42 expertos, de los que 31 eran mujeres y 11 eran hombres. El grueso de los expertos, 34 personas, están dentro del rango de edad de los 18 a los 35 años y sólo 8 eran mayor de 35 años. Todos los participantes eran hablantes nativos de español, mayores de edad y profesores.

Para realizar la evaluación diseñamos un cuestionario usando la herramienta Google Form[4], que permite crear formularios online y recopilar las respuestas a las preguntas planteadas. A los participantes se les presentaron 15 pares de oraciones, original y simplificada por el sistema, con 34 expresiones numéricas de distintos tipos (numerales, partitivos, porcentajes y monetarias) y con una media de 33,5 palabras por oración y una media de 2,26 expresiones numéricas por oración. Las respuestas son dicotómicas, sólo tienes dos opciones, sí o no, y para cada par de frases, se preguntaban tres cosas:

1. Si la oración simplificada preservaba el significado de la original.

[4] http://bit.ly/1wMwCwZ



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Preserva el significado	78,57%	78,57%	76,19%	85,71%	83,33%	80,95%	90,47%	61,90%	85,71%	85,71%	61,90%	90,47%	92,85%	85,71%	85,71%
Gramaticalmente correcta	95,23%	54,76%	92,85%	92,85%	73,80%	69,04%	92,85%	90,47%	85,71%	83,33%	64,28%	76,19%	59,52%	78,57%	76,19%
Exp.Num. Bien simplificadas	78,57%	69,04%	66,66%	71,42%	61,90%	66,66%	83,33%	47,61%	88,10%	80,95%	52,38%	78,57%	90,47%	80,95%	73,80%

Figura 5.10: Datos recogidos en el cuestionario con expertos en español para evaluar la salida del sistema de simplificación de expresiones numéricas

2. Si la oración simplificada era gramaticalmente correcta.

3. Si las expresiones numéricas estaban bien simplificadas.

Analizando los datos recogidos en el cuestionario, los resultados muestran que los participantes consideran que la versión simplificada de las oraciones que genera el sistema preservan el significado en comparación con las oraciones originales con una media de 81,58 % y una desviación estándar de 9,24 %. Además, los participantes creen que la oración con las expresiones numéricas simplificadas es gramaticalmente correcta con una media de 79,04 % y una desviación estándar de 12,98 %. Finalmente, consideran que las expresiones numéricas fueron simplificadas correctamente con una media de 72,69 % y una desviación estándar de 12,3 %.

Para cada proporción de la muestra se calcula la inferencia estadística construyendo el intervalo de confianza (IC) al 95 %. La amplitud del intervalo obtenida depende del porcentaje de la muestra y del tamaño de la misma que lo sustenta. Como podemos ver para la primera pregunta se construyó un IC [78,6 %, 84,6 %] con un error estándar del 1,5 %. Mientras que el valor del 50 % no esté cubierto por el IC, estos datos confirman que el número de respuestas afirmativas es estadísticamente mayor que el número de respuestas negativas, y sabemos entonces que la oración simplificada conserva el significado de la oración original según la opinión de los participantes. Para la segunda pregunta hemos construido un IC [75,9 %, 82,2 %] con un error estándar del 1,6 %. Es el mismo caso y estos datos confirman que la oración simplificada es gramaticalmente correcta. Para la última pregunta hemos construido un IC [69,2 %, 76,2 %] con un error estándar de 1,8 %.

Aquí también podemos ver que el número de respuestas afirmativas es estadísticamente mayor que el número de respuestas negativas, por lo que los datos confirman que las expresiones numéricas fueron simplificadas correctamente de acuerdo con la opinión de los participantes. En la Figura 5.10 podemos ver los datos para cada una de las 15 oraciones del cuestionario.

En cuanto a decidir si las expresiones numéricas están bien simplificadas por el sistema, confiamos en su criterio de evaluación, debido a que los profesores están formados con conocimientos de pedagogía y adaptación curricular. De ahí que sean el tipo de expertos seleccionados para evaluar la salida de nuestro sistema.

5.5. Comparación de los sistemas de simplificación de expresiones numéricas implementados

Los dos sistemas implementados llevan a cabo la tarea de simplificar automáticamente expresiones numéricas en un texto. Cada uno de ellos tiene sus características y ambos comparten cosas en común. A continuación hacemos una comparativa entre ambos sistemas.

El sistema en inglés simplifica sólo las expresiones numéricas en porcentajes. Aunque sólo trata un tipo de expresiones numéricas, el sistema ofrece la posibilidad de adaptar dichas expresiones a distintos niveles de dificultad. Además, por las herramientas y recursos que utiliza, el sistema utiliza expresiones no numéricas para los valores extremos de la proporción de entrada y es capaz de generar fracciones candidatas para simplificar las fracciones originales que hay en el texto.

El sistema en español trata un rango más amplio de expresiones numéricas como son los numerales, las expresiones monetarias, los porcentajes y las fracciones, pero no identifica distintos niveles de dificultad en el proceso de simplificación. Además, no es capaz de generar expresiones no numéricas para simplificar determinadas expresiones y no adapta las fracciones con fracciones equivalentes en el proceso de simplificación.

Ambos sistemas implementan una regla para transformar las expresiones representadas en letras por su correspondiente versión en dígitos. Las expresiones numéricas representadas en ratios no son tratadas en ninguno de los dos sistemas. Los porcentajes siempre son redondeados al valor más próximo, subsanando esta pérdida de precisión con el uso de modificadores. Los dos sistemas tienen un conjunto de reglas para determinar qué modificador utilizar en la versión simplificada de la expresión numérica que están tratando. En general, si la expresión numérica original tiene modificador, entonces éste se mantiene. Si no tiene y hay pérdida de precisión, entonces se añade un modificador. Si no hay pérdida de precisión, no se añade modificador.

Como podemos ver, a partir del estudio empírico que se realizó siguien-

do diferentes metodologías, se han podido definir e implementar reglas de simplificación que permiten llevar a cabo la simplificación automática de expresiones numéricas presentes en los textos. En este trabajo hemos presentado dos sistemas automáticos de simplificación, uno para el inglés y otro para el español. Pero con nuestra propuesta de trabajo se podría abordar la tarea de simplificación para cualquier otro idioma.

Resumen y conclusiones

En este capítulo hemos presentado dos sistemas de simplificación de expresiones numéricas, uno para inglés y otro para español. Ambos sistemas han sido implementados a partir del modelo genérico especificado en el capítulo anterior y definiendo reglas a partir de las estrategias identificadas en el estudio empírico llevado a cabo. Además han sido evaluados por expertos que han examinado la salida generada por los sistemas para las expresiones numéricas simplificadas.

Para el sistema de simplificación en inglés se les presentó la salida del sistema en dos niveles. Para el nivel donde las expresiones numéricas eran simplificadas usando fracciones, en un 54 % de los casos los expertos estaban de acuerdo con la salida generada por el sistema. En el nivel donde los porcentajes eran usados sin decimales para simplificar las expresiones originales, los expertos estaban de acuerdo en un 79 % de los casos. Son resultados positivos que demuestran que las reglas aplicadas para simplificar en inglés tienen una alta aceptación.

Para el sistema en español se midió primero la cobertura de las reglas definidas para la simplificación de las expresiones numéricas con casi un 84 % de precisión. Además, se hizo una evaluación con expertos que nos permitió analizar las simplificaciones hechas por el sistema. Las respuestas mostraron que casi un 82 % de los expertos estaban de acuerdo con que la versión simplificada preservaba el significado de la oración original, un 79 % estaba de acuerdo con que la versión simplificada era gramaticalmente correcta y casi un 73 % consideraron que las expresiones numéricas estaban bien simplificadas. Estos resultados apoyan las estrategias de simplificación implementadas en el sistema y permiten seguir trabajando en futuras mejoras del sistema.

Capítulo 6

Discusión

En las siguientes secciones presentamos primero una discusión del planteamiento y desarrollo del trabajo. A continuación veremos el modelo genérico presentado como una abstracción de la práctica existente, junto con la discusión de las metodologías seguidas para la identificación experimental de las estrategias de simplificación de expresiones numéricas. Finalmente, discutiremos los sistemas implementados.

6.1. Discusión del planteamiento y desarrollo del trabajo

Dentro de las múltiples opciones que se abordan en el campo de la simplificación de textos, había que decidir centrar el foco de atención en un tipo concreto y afrontar el problema planteando una solución. Debido a que hasta ahora la información numérica de los textos casi no había sido tratada en el área de la simplificación, el trabajo de esta tesis se ha centrado en el proceso de simplificación de las expresiones numéricas presentes en un texto. A partir de ahora cuando se quiera abordar el tema de simplificación de expresiones numéricas se podrán contemplar los aspectos identificados que influyen en el proceso automático de simplificación.

El trabajo de esta tesis presenta un modelo para la simplificación automática de expresiones numéricas y la implementación de dos sistemas computacionales que realizan la simplificación para textos en inglés y para textos en español.

El modelo genérico propuesto presentado en el capítulo 4 cubre una serie de variables (tipo de texto, lenguaje, nivel de dificultad y usuario final) que no son cubiertas todas ellas a la vez en los sistemas implementados. Cada sistema presentado en esta tesis cubre distintos aspectos. Por ejemplo, el sistema de simplificación de inglés cubre el nivel de dificultad en la adaptación de un tipo de texto que son las noticias, mientras que el usuario final es una

variable considerada sólo para el caso del estudio en español con personas con dislexia. Somos conscientes de que el caso ideal sería tener un sistema que cubriera todas las variables de la mejor manera posible, es decir, todos los niveles de dificultad definidos, el usuario final, el tipo de texto y el lenguaje. Esta aproximación queda como trabajo futuro de esta tesis.

Las herramientas y recursos necesarias para la implementación del modelo genérico fueron presentadas en el capítulo 3 y en concreto las herramientas específicas para la simplificación de expresiones numéricas. La primera decisión que hay que tomar es el conjunto de textos con el que vamos a trabajar, de ahí que la elección del corpus sea tan importante. En nuestro trabajo, tanto para el caso del inglés como para el español, contamos con los corpus de noticias utilizados en distintos proyectos de investigación que nos brindaron la oportunidad de utilizarlos como material.

Para poder definir e implementar computacionalmente las reglas que se deben aplicar en el proceso de simplificación automática de expresiones numéricas, se llevó a cabo una identificación experimental con expertos como mostramos en el capítulo 4. El procedimiento a seguir parte de plantear unas intuiciones de partida que queremos validar con los expertos, se lleva a cabo una selección del material necesario, se realiza el diseño del estudio siguiendo distintas metodologías y finalmente se implementa el estudio, para después analizar los datos recogidos y poder validar o no nuestras intuiciones de partida.

La decisión de usar unas herramientas u otras en cada etapa del modelo determina detalles del diseño del estudio y del sistema que se quiera implementar (capítulo 5). Pero una vez que las estrategias de simplificación están identificadas, el proceso de definirlas e implementarlas para el lenguaje concreto es un proceso de instanciación de la metodología identificada y presentada. Estas decisiones se toman dependiendo del lenguaje con el que se está trabajando y teniendo en mente la finalidad del sistema, ya que se pretende que sea una herramienta de ayuda para las personas que tienen que adaptar textos para personas con dificultad en la lectura y comprensión de la información a la que están accediendo.

6.2. El modelo genérico como una abstracción de la práctica existente

El modelo genérico para el proceso de simplificación de textos presentado en esta tesis (sección 4.1) se explica como una abstracción cuyo objetivo es cubrir un número de procedimientos que son seguidos por los sistemas de simplificación ya existentes. El lenguaje con el que estamos trabajando determina las herramientas que se utilizan en cada etapa para llevar a cabo el proceso de simplificación de expresiones numéricas. Prestamos especial atención a la etapa 2 del modelo donde en la descomposición del texto se

va a realizar la identificación y anotación de las expresiones numéricas, y a la etapa 3 de simplificación del texto, donde se ejecutan las reglas implementadas a partir de lo aprendido en la identificación experimental para determinar las estrategias de simplificación y el uso de modificadores llevada a cabo previamente.

Por ejemplo, el sistema presentado por Carroll et al. (1998) ayuda a personas afásicas simplificando automáticamente noticias en inglés disponibles en Internet. El sistema se puede dividir en dos componentes principales: un componente para el análisis, que proporciona el etiquetado léxico, un análisis morfológico y un análisis sintáctico; y un componente para la simplificación, que adapta la salida del analizador para facilitar la lectura a las personas afásicas utilizando transformaciones léxicas y sintácticas.

En términos del modelo que se describe en esta tesis, el componente para el análisis correspondería a la etapa 1 de Análisis del Texto. Las transformaciones léxicas corresponderían a una instanciación particular de la etapa 2 - Descomposición del Texto - para determinar las palabras difíciles como unidades objetivo, y una instanciación particular de la etapa 3 - Simplificación del Texto - en la que se aplica las sustituciones de estas palabras difíciles para dar lugar a alternativas más simples. Las transformaciones sintácticas corresponderían a una instanciación particular de la etapa 2 - Descomposición del Texto - para producir particulares construcciones sintácticas como unidades objetivo, y una instanciación particular de la etapa 3 - Simplificación del Texto - donde se aplican las transformaciones a estas construcciones sintácticas para dar lugar a formulaciones más simples. Para ambas instancias, un proceso final de la reconstrucción de la versión completa del texto simplificado corresponde a la etapa 4 - Regeneración del Texto - tal y como se describe en la sección 4.1.4 del modelo genérico.

Analizamos otros sistemas de manera similar. En el proyecto *Simplext* (Saggion et al., 2011), donde se simplifican noticias de prensa en español para personas con dificultades cognitivas, el texto se analiza usando las herramientas FreeLing (Padró et al., 2010) y GATE (Cunningham et al., 2002), análisis que se corresponde con la etapa 1 - Análisis del Texto - de nuestro modelo genérico. La aplicación posterior de las transformaciones léxicas y sintácticas puede considerarse como instanciaciones de las etapas 2 y 3 - Descomposición del Texto y Simplificación del Texto, llevando finalmente a cabo una instanciación de la etapa 4 - Regeneración del Texto - para generar la versión final simplificada.

El proyecto *PorSimples* (Specia, 2010) desarrolló herramientas para el portugués brasileño y tiene como objetivo el desarrollo de tecnologías para que el acceso a la información sea más fácil para las personas de baja alfabetización. Esta propuesta establece que la simplificación del texto se puede subdividir en simplificación sintáctica, simplificación léxica, resumen automático y otras técnicas. Esta proliferación de operaciones puede ser vis-

ta como la integración de varias instancias de nuestro modelo genérico, con diferentes tipos de operaciones de simplificación que se aplica a diferentes niveles de granularidad de la descomposición (resumen del texto completo, reescritura sintáctica a nivel de construcciones sintácticas, sustitución de palabras a nivel de términos léxicos).

Tras la revisión de estos tres sistemas pertenecientes a la simplificación de textos en varios idiomas, hemos visto que pueden ser descritos en términos de nuestro modelo genérico de simplificación de texto, y esto se puede tomar como un indicador de un cierto grado de generalidad que puede ayudar a mejorar la comparabilidad entre los distintos sistemas. En cada caso particular, el idioma con el que se trabaja, las herramientas que se van a utilizar, el tipo de texto y el usuario final para el que se está simplificando, tienen que ser definidos. Sobre éstos, cada sistema aplica su análisis y, dependiendo del objetivo del sistema, se definen las transformaciones de simplificación específicas para aplicarlas y generar la versión final simplificada del texto original.

Además, desde un punto de vista más abstracto, podemos ver que el modelo de proceso para sistemas de simplificación de textos propuesto por Siddharthan (2002) y el modelo genérico de simplificación que presentamos en esta tesis siguen la misma idea en el proceso de simplificación para generar la versión simplificada de un texto original. Vamos a ver las similitudes y diferencias entre ambas propuestas.

La arquitectura de Siddharthan permite que cada componente sea desarrollado y evaluado de manera independiente. En su propuesta hace especial énfasis en aspectos a nivel de discurso en las operaciones de simplificación sintácticas como es generar expresiones de referencia, decidir determinantes, decidir el orden de la oración y preservar las estructuras retóricas y anafóricas.

Siddharthan en su trabajo propuso una arquitectura formada por tres fases: análisis, transformación y regeneración. La primera fase se encarga de generar la representación estructural de la oración, a nivel de análisis sintáctico y etiquetado de categorías gramaticales. La segunda fase usa reglas de transformación para generar texto plano a partir de la estructura conseguida por el estado anterior. Y la tercera y última fase es la encargada de realizar las simplificaciones sintácticas que se contemplan en cada caso.

En cambio, el modelo genérico presentado en esta tesis consta de cuatro fases o etapas: análisis, descomposición del texto, simplificación del texto y regeneración del texto. Aunque nuestra última etapa se llame igual que la tercera fase de la arquitectura de Siddharthan, no coinciden en funcionalidad, ya que se realizan operaciones diferentes en un estado y en otra. La primera etapa de nuestro modelo es la encargada del análisis del texto, a nivel de estructuras sintácticas y etiquetado de categorías gramaticales. En la segunda etapa se descompone el texto identificando las unidades lingüís-

ticas que van a ser simplificadas. En la tercera etapa es donde se aplican las reglas de simplificación para generar las versiones simplificadas de las unidades identificadas. Y finalmente, la etapa de regeneración se encarga de recomponer el texto con las versiones simplificadas de las unidades tratadas, para generar la versión simplificada del mismo.

Comparando ambas aproximaciones, podemos ver que la etapa inicial de análisis es común a ambas, ya que dado un texto original que se quiere simplificar, el primer paso a seguir es realizar un análisis del mismo para utilizar esta información en las etapas siguientes. La siguiente etapa es diferente para ambos casos, ya que en la arquitectura de Siddharthan consiste en generar texto plano a partir de las estructuras obtenidas en el análisis, mientras que en el modelo que se propone en este trabajo la segunda etapa se corresponde con la identificación de las unidades lingüísticas que se van a simplificar. La tercera etapa, que es donde propiamente se llevan a cabo las transformaciones de simplificación, en la arquitectura de Siddharthan se llama regeneración y en nuestro modelo se llama simplificación de texto. La idea es común en ambos casos, ya que lo que se aplican son reglas de transformación centradas en la simplificación que se quiera realizar. Además, el modelo propuesto en este trabajo contempla una etapa más para recomponer el texto con las unidades simplificadas y generar la versión simplificada.

6.3. Identificación experimental realizada

La falta de pautas definidas en el proceso de simplificación de expresiones numéricas nos llevó a realizar una identificación experimental con expertos, para que analizando los datos recogidos pudiéramos definir e implementar las reglas computacionales que nos permiten realizar la simplificación automática de expresiones numéricas de un texto. Somos conscientes de que no existen reglas para determinar cuándo una expresión numérica debería ser simplificada o no. De ahí que nuestra propuesta vaya dirigida a expertos que trabajan en el ámbito de adaptación de contenidos para que les ayude a realizar el proceso de simplificación según su criterio. Con su ayuda, se especificarán los tipos de expresiones numéricas que se quieren simplificar y así, en la etapa de descomposición del texto, se identificarán las unidades correspondientes que se quieran simplificar.

En el caso del sistema implementado para el inglés se decidió simplificar las expresiones numéricas presentadas en porcentajes, y a partir de esa decisión se realizó la identificación experimental de las estrategias que los humanos usaban para simplificar los porcentajes que aparecían en un texto. Para el caso del sistema implementado para el español, se amplió el rango de expresiones numéricas, dando lugar a poder simplificar numerales, expresiones monetarias, numerales partitivos, fracciones y porcentajes.

Lo complicado en el proceso de identificación de las estrategias de sim-

plificación es contar con un grupo de expertos que nos ayuden a realizar el proceso de simplificación manual para luego identificar patrones de simplificación que pudiésemos implementar. En este trabajo contamos con la ayuda de profesores de matemáticas y de personas que trabajan en el campo de adaptación de contenidos para personas con discapacidad y consideramos que era un grupo valioso por su formación y dedicación.

Los cuestionarios que realizamos a los expertos fueron diseñados e implementados usando herramientas online de diseño de formularios que nos permitieron tener un mayor alcance para obtener el mayor número de participantes y tener los datos digitalizados para su correspondiente análisis. A la hora de realizar este tipo de estudios, el diseño del cuestionario es muy importante para que cubra las expectativas que se desean, para que no tenga mucho ruido que contamine los datos y para que los resultados que obtengamos sean útiles para su futura implementación. Somos conscientes de que el presentar las frases con sus correspondientes expresiones numéricas fuera del contexto del texto a veces puede resultar extraño, pero viendo el estudio que realizamos con el análisis de la simplificación hecha a mano por expertos en el proyecto de *Simplext* (sección 3.1), en la que en muchas ocasiones las expresiones numéricas en lugar de ser simplificadas, eran eliminadas, nos planteamos realizar los estudios forzando a que las personas tuvieran que simplificar, obligatoriamente, las expresiones que se les mostraban.

6.4. Sistemas de simplificación de expresiones numéricas implementados

Los sistemas centrados en simplificar expresiones numéricas que presentamos en esta tesis siguen el modelo genérico de simplificación eligiendo como unidades de simplificación las expresiones numéricas e implementando las reglas de transformación identificadas anteriormente para generar su versión simplificada y recomponer el texto original con las expresiones numéricas simplificadas.

Para cada sistema hay que decidir las herramientas usadas en cada etapa del modelo. La implementación de los sistemas presentados en este trabajo se realizó decidiendo entre unas herramientas u otras dependiendo de la cobertura de las mismas y de los objetivos planteados para cada sistema. La decisión de qué analizador utilizar viene ligada al tipo de análisis que queremos y al detalle de la información que necesiten las reglas definidas para el tratamiento de la información. De ahí que para el caso del inglés se descartó el *Stanford Parser*, ya que no identificada las expresiones numéricas con suficiente información para las reglas que queríamos implementar, y se decidió usar el *parser* específico diseñado por la Dr. Sandra Williams. Este *parser* nos permitió etiquetar los modificadores como parte de la expresión numérica, cosa necesaria para las estrategias que habíamos definido. Para el

caso del español, el analizador elegido fue FreeLing ya que nos proporcionaba un análisis detallado de las expresiones numéricas, información que luego utilizamos para nuestras gramáticas y reglas de simplificación.

Como herramientas específicas del trabajo, señalamos para el caso del inglés el uso del *programa de aproximación de proporciones*. El mayor inconveniente que tiene esta herramienta es que sólo trabaja con proporciones normalizadas (0,0 a 1,0), por lo que no se puede extender su uso para grandes números y esto es una restricción a la hora de realizar la simplificación de información numérica en un texto. Para el caso del español, la herramienta escogida para el tratamiento de expresiones numéricas fue JAPE (*Java Annotation Patterns Engine*), que nos permitió tratar un conjunto más amplio de expresiones numéricas y dar mayor cobertura a la hora de simplificar información numérica en español. La mayor atadura que tiene usar JAPE es su dependencia con el entorno de desarrollo GATE, pero a su vez da mucha flexibilidad para poder identificar y anotar todo tipo de información en el texto. Con la ayuda de GATE se definieron un par de *plugings* que nos permitieron realizar el proceso de simplificación de una manera gráfica para visualizar las identificaciones, las anotaciones y las simplificaciones realizadas a lo largo del proceso.

Debido a la falta de métricas automáticas para verificar si la versión simplificada de una expresión numérica es más fácil de entender que la original, recurrimos de nuevo a nuestros expertos para que evaluaran la salida automática generada por los sistemas. Es importante medir hasta qué punto la salida generada sea gramaticalmente correcta, preserv razonablemente el significado de la versión original y la simplificación de las expresiones numéricas sea correcta. De ahí nuestra decisión de contar de nuevo con nuestros expertos recurriendo a encuestas online para recoger su opinión y analizar los datos recogidos. Es complicado evaluar este tipo de sistemas en el que la opinión de los expertos no deja de estar sujeta a una opinión subjetiva y personal de cada persona, pero con su ayuda se realizó la validación de los sistemas de simplificación implementados.

Así, hemos comprobado nuestra suposición de partida de que en el proceso de simplificación de expresiones numéricas, a la vez que se disminuye la dificultad numérica de comprender la expresión matemática en sí, se aumenta la complejidad sintáctica de la expresión. La mayoría de las estrategias de simplificación aplicadas en el proceso automático realizado hacen que la expresión numérica original sea sustituida por una expresión formada por más componentes sintácticos que la original, como es el caso al añadir un modificador cuando la cantidad numérica es redondeada, o cambiar la representación matemática generando una expresión más compleja sintácticamente que la original, pero matemáticamente más sencilla. Veamos los siguientes casos como ejemplos:

- La expresión numérica original *48,6 %* se simplificaría por la expre-

sión *casi el 50 %*. Podemos ver que debido a la pérdida de precisión realizada por el redondeo de la cantidad original, se ha añadido un modificador. Por lo tanto, a nivel sintáctico, la expresión simplificada es más compleja a pesar de que la dificultad matemática es menor.

- La expresión numérica original *500 ml* se simplificaría por la expresión *1/2 l*. Se puede ver que en este proceso de simplificación se ha optado por hacer un cambio de representación matemática, de numeral a fracción, produciéndose además un cambio de unidades, debido a que la frecuencia de uso aquí juega un papel fundamental. Es muy importante que la representación matemática y las unidades correspondientes sean frecuentemente usadas y conocidas para que ayude a su comprensión.

La evaluación realizada con expertos nos permitió analizar la salida que genera nuestros sistemas, planteando líneas de trabajo futuro de mejora de los sistemas, definiendo otro tipo de estrategias de simplificación y mejorando las evaluaciones realizadas.

6.4.1. Evaluación del sistema de español en un *pipeline* externo

Como un caso especial de evaluación de nuestro sistema de simplificación de expresiones numéricas en español, se realizó la integración de nuestro sistema como un módulo más de la parte léxica de un *pipeline* más completo de simplificación automática de textos en español (Drndarevic et al., 2013). El pipeline consiste en un módulo basado en reglas de transformación léxica y un módulo para la simplificación sintáctica. Los dos componentes han sido evaluados por separado y en conjunto, para determinar el nivel de simplificación, de preservación del significado y gramaticalidad.

Para evaluar el nivel de legibilidad antes y después del proceso de simplificación, se aplican fórmulas de legibilidad para el español (Spaulding, 1951; Anula, 2007) sobre cuatro conjuntos elegidos aleatoriamente de artículos de prensa: el texto original, la salida obtenida después de las transformaciones léxicas, la salida después de la simplificación sintáctica, y la salida de ambos módulos conjuntamente. Para evaluar si la salida simplificada era gramaticalmente correcta y semánticamente adecuada, se realizó una encuesta con personas presentándolas pares de oraciones, original y simplificada.

Los resultados indican que los componentes del sistema (módulo de simplificación sintáctica y transformaciones léxicas basados en reglas) producen una salida más simple en comparación con el original, y que la combinación de los dos logra un mayor grado de simplificación que cualquiera de los elementos individualmente.

En cuanto a la precisión lingüística de la salida, el pipeline fue valorado positivamente por los participantes. El 60 % de los mismos considera que las

oraciones simplificadas son gramaticalmente correctas, mientras que alrededor del 70 % de ellos estuvo de acuerdo en el hecho de que el significado se ha conservado bastante bien en el proceso de simplificación. El análisis cualitativo de los resultados reveló que la mayoría de los errores comunes que resultan en la falta de gramaticalidad fueron malas estrategias de coordinar las estructuras en la etapa de simplificación sintáctica y la falta de tratamiento del contexto a la hora de aplicar las transformaciones léxicas.

6.5. Interpretación de las expresiones numéricas

La interpretación de las expresiones numéricas es una tarea compleja en la que se involucra tanto un ingrediente de procesamiento matemático (para tratar características como la cantidad y el tipo de representación matemática) y un ingrediente de procesamiento de lectura (para tratar con el amplio contexto en el que la expresión ocurre, pero también para interpretar cualquier modificador que las acompañe y para identificar la información numérica expresada con expresiones no numéricas). En este sentido, las personas con diferentes niveles de competencia en los correspondientes procesos pueden encontrar algunas formas de expresar la información numérica más difíciles que otras.

A los buenos lectores con pobres habilidades matemáticas, por ejemplo, les resulta más fácil de entender la información numérica cuando es expresada, si es posible, en términos de expresiones no numéricas. Por el contrario, los lectores con menor habilidad lectora pero con buenas habilidades matemáticas prefieren formulaciones complejas matemáticas con menos uso de modificadores y expresiones textuales. Estas características deben tenerse en cuenta en cualquier proceso de simplificación, donde una descripción adecuada de las capacidades del usuario final para el que se está simplificando debe ser considerada como dato de entrada al definir e implementar las operaciones de simplificación que se van a aplicar en el proceso de simplificación.

En relación con los aspectos cognitivos, los diversos factores que intervienen en los procesos de lectura (presentados en la sección 2.1.2) deberían jugar un papel crucial. En particular, los factores de percepción afectarán a la capacidad del usuario final para percibir la entrada, y esto debe ser tenido en cuenta a la hora de seleccionar entre las formas alternativas de presentación. Factores psicolingüísticos y metalingüísticos desempeñan un papel fundamental en los cambios y decisiones, léxicas y sintácticas, que pueden estar implicados en las transformaciones aplicadas durante el proceso de simplificación.

Con el fin de perfeccionar los dos principios, inmediatez e interactividad, que explican el proceso de lectura de un texto, éstos deben ser considerados como principios opuestos, y un equilibrio entre la economía del esfuerzo seguido por el primero y la precisión con respecto al contenido por el otro, debe

ser alcanzado. Esto es relevante para la simplificación, ya que las expresiones más simples que favorecen la inmediatez pueden conducir a errores en la comprensión. Por el contrario, expresiones más complejas pueden producir una impresión inicial de desconcierto, pero cuando este desconcierto conduce a una mayor interacción con la formulación, el efecto general puede ser un aumento en la precisión de la comprensión.

Muchos de los factores que entran en juego en los procesos de lectura y razonamiento matemático tienen una gran importancia a la hora de llevar a cabo el proceso de simplificación de textos, en concreto en la simplificación de expresiones numéricas. Los aspectos cognitivos tienen que ser considerados para una adecuada identificación de las estrategias de simplificación que se quieran automatizar, mejorando así la cobertura de las reglas implementadas y consiguiendo una versión simplificada más cercana al usuario final.

Resumen y conclusiones

En este capítulo se recoge la discusión de los distintos aspectos que se han presentado en los capítulos anteriores de esta tesis. Partiendo de la discusión del planteamiento y desarrollo del trabajo, hemos ido viendo las distintas decisiones tomadas a nivel de herramientas y recursos necesarias para la implementación final de los sistemas de simplificación. Después se discute el modelo genérico como una abstracción de la práctica existente y continuamos con la discusión de la identificación experimental y de los sistemas implementados para la simplificación de expresiones numéricas.

En el siguiente capítulo presentamos las conclusiones y las líneas de trabajo futuro que planteamos como continuación del trabajo realizado hasta ahora.

Capítulo 7

Conclusiones y Trabajo Futuro

En el trabajo de esta tesis hemos presentado un modelo genérico para la simplificación de textos, y en particular hemos descrito e implementado dos sistemas de simplificación de expresiones numéricas en inglés y en español. Para definir las reglas a implementar se realizó un estudio empírico para identificar las estrategias de simplificación que usamos las personas cuando tenemos que simplificar las expresiones numéricas que hay en un texto. Una vez desarrollados los sistemas, se llevó a cabo una evaluación con expertos para validar la salida automática generada y así plantear futuras líneas de mejora del trabajo desarrollado.

Este capítulo final de la tesis recoge las conclusiones y el trabajo futuro del trabajo expuesto. En la sección 7.1 se hace un repaso de las principales conclusiones. La sección 7.2 muestra las líneas de trabajo futuro que se desprenden de esta tesis.

7.1. Conclusiones

Al finalizar el trabajo de esta tesis hemos recogido algunas reflexiones que nos permiten encuadrar el trabajo realizado a partir de las necesidades de la sociedad en la que vivimos y analizando la aportación realizada con nuestro trabajo de investigación. Primero presentamos la importancia de la simplificación de textos en la educación, revisando la necesidad de atender a la diversidad de la sociedad ante la que nos encontramos. Y después, presentamos unas conclusiones generales de nuestro trabajo a partir del modelo genérico y de los sistemas implementados.

7.1.1. La simplificación automática de textos

Los cambios en la *Sociedad de las Tecnologías de la Información* nos llevan a considerar los cambios en el tratamiento y procesamiento de la información. Por ejemplo, la simplificación de textos manual no puede hacer

frente al proceso de adaptación de la gran cantidad de contenidos que se generan para diversas audiencias, ya que requiere una gran cantidad de tiempo y esfuerzo. Esta realidad nos lleva a tomar en consideración soluciones tecnológicas que nos ayuden a mejorar el acceso a la información para personas con dificultades especiales.

La alfabetización entendida como la comunicación escrita involucra procesos cognitivos de lectura que exigen esfuerzo y presentan dificultades para las personas con problemas cognitivos. El procesamiento de la información numérica juega un papel fundamental en esta alfabetización debido a que las expresiones numéricas se presentan en diferentes contextos, tales como noticias, recetas, facturas, etc. Nuestra principal motivación para automatizar el proceso de simplificación de expresiones numéricas es la dificultad que algunas personas pueden tener para entender este tipo de información en un texto.

En este trabajo se ha definido un modelo genérico para llevar a cabo la simplificación automática de textos, identificando las variables importantes que tienen que ser consideradas en el proceso. Nos centramos en el tratamiento de la información numérica como un caso especial de estudio y validamos nuestro modelo instanciándolo en dos sistemas reales de simplificación centrada en expresiones numéricas de textos en inglés y en español. Ambos sistemas fueron evaluados por expertos en el área que nos han permitido analizar los resultados recogidos y plantearnos futuras mejoras. Además, para el caso del español, presentamos un caso de estudio real con personas con dislexia que realizamos para comprobar nuestras hipótesis de trabajo y que nos han permitido ver de cerca la realidad de un colectivo concreto, y conocer así las estrategias específicas que este grupo de usuarios necesita.

En el modelo genérico hemos presentado diferentes etapas en el proceso de simplificación automática de textos. Consideramos una etapa de simplificación especial donde gran cantidad de diferentes transformaciones puede ser consideradas, dependiendo del idioma, del nivel de dificultad, del tipo del texto original y del usuario final para el que se está adaptando el texto. Estas transformaciones se pueden aplicar a nivel de oración o de palabra, pero podrían ser consideradas en un nivel superior, cómo por ejemplo el párrafo. Posibles simplificaciones incluyen diferentes tipos de operaciones tales como resúmenes, paráfrasis, adición o eliminación de información, evitar la metáfora, el sarcasmo o la ironía, etc. Para automatizar este tipo de operaciones, otro tipo de variables a nivel pragmático del lenguaje como el contexto o la semántica, deben ser consideradas.

Con la intención de centrarnos en la simplificación de la información numérica en los textos, redefinimos la etapa de simplificación del texto en el modelo con el enfoque en las expresiones numéricas. El lenguaje del texto original influye en el proceso de simplificación, ya que afecta a la identificación y anotación de las expresiones numéricas en la etapa de descomposición

del texto. Ambas etapas son dependientes de la información proporcionada por la etapa previa del análisis de texto. Este análisis es la base para el reconocimiento de expresiones numéricas en el texto, su posterior anotación de la información como modificadores, unidades o cantidades.

Para obtener las reglas de simplificación que se van a automatizar en cada sistema, se realizó una identificación experimental de las estrategias de simplificación que utilizan los humanos a la hora de realizar simplificaciones de expresiones numéricas. El proceso de identificación se realizó con expertos, y de este modo tenemos información sobre las diferentes transformaciones que se pueden aplicar para simplificar y el uso que hacen de los modificadores a la hora de generar la versión simplificada de las expresiones numéricas. Nuestros estudios muestran que el valor de la proporción en la expresión numérica influye en la estrategia y que la forma matemática final y el uso de modificadores son factores importantes en el proceso de simplificación de las expresiones numéricas.

Es importante destacar que la adaptación de contenidos es necesaria para cubrir los distintos niveles que hay en el aula, a nivel educativo, para que la información sea accesible al mayor número de personas. La finalidad del trabajo presentado es ayudar a los expertos a adaptar los contenidos y ser capaces de agilizar este proceso para que la *Sociedad de las Tecnologías de la Información* sea una realidad para todos los individuos que forman parte de ella.

7.1.2. La importancia de la simplificación de textos en la educación

Europa está cambiando a un ritmo comparable al de la revolución industrial. Por un lado, las tecnologías digitales están transformando todos los aspectos de la vida de las personas. Por otro lado, el comercio, los viajes y la comunicación a nivel global están ampliando los horizontes culturales y cambiando las pautas de competencia de las economías. La vida ofrece, en la actualidad, mejores oportunidades y opciones, pero también entraña mayores riesgos e incertidumbres. Las personas tienen la libertad de adoptar estilos de vida diferentes, pero también la responsabilidad de dar forma a sus propias vidas.

Hoy en día se da la paradoja de que existen más ciudadanos que prolongan su educación y su formación, pero, al mismo tiempo, se está aumentando la desigualdad entre los que gozan de una cualificación suficiente para mantenerse activos en el mercado de trabajo y los que quedan irremediablemente desplazados. Este fenómeno va a cambiar la composición de la población activa y las pautas de demanda de servicios sociales, sanitarios y educativos. Las sociedades europeas se están convirtiendo en mosaicos interculturales, por lo que esta diversidad encierra un gran potencial para la creatividad y la innovación en todos los ámbitos de la vida.

El término *Sociedad de las tecnologías de la información* va siendo cada vez más utilizado y refleja el hecho de que la sociedad está siendo cada vez más tecnológica y que pretende hacer frente a las cuestiones económicas, culturales, sociales y laborales de la actualidad. El aprendizaje permanente, arraigado en la realidad de nuestra sociedad, está estrechamente relacionado con el mundo del trabajo, el mercado, su cambio y su ritmo evolutivo.

Además, la tecnología tiene una influencia decisiva en la educación de los individuos. En el mundo actual hay personas que, por diversas razones, nunca fueron a la escuela o no pudieron lograr resultados a largo de su educación. Adultos de todas las edades tienen que adaptarse constantemente, a través del aprendizaje, a las circunstancias cambiantes de la vida. Es por tanto necesario establecer mecanismos que ofrezcan posibilidades educativas atractivas para estas personas, que atiendan a sus necesidades específicas y mecanismos de adaptación de los materiales utilizados en el curso para su educación. Es necesario superar los obstáculos que impiden la educación atendiendo a las necesidades de la sociedad moderna, en particular, a las necesidades de los individuos para adaptarse a los cambios en todas las etapas de su vida.

En el proceso de superación de estos obstáculos, es importante también tener en cuenta las dificultades a las que cada individuo se enfrenta, asumiendo que independientemente de las deficiencias individuales, todo el mundo tiene en común una dificultad de aprendizaje. Utilizamos el término dificultades de aprendizaje para identificar esas deficiencias en aspectos instrumentales de aprendizaje, en particular deficiencias lingüísticas, lógicas y matemáticas, que impiden el desarrollo normal de los contenidos curriculares en los diferentes campos.

Proporcionar la educación apropiada a cada uno de los estudiantes se basa en el principio de inclusión, siendo la única manera de garantizar su crecimiento, promover la igualdad y contribuir a una mayor cohesión social. La atención a la diversidad es una necesidad que se aplica a todas las etapas educativas y a todos los alumnos. Es decir, la diversidad de los estudiantes debe ser abordada como un principio, y no como una situación relativa a las necesidades de un grupo reducido de estudiantes.

7.2. Trabajo Futuro

El trabajo descrito en esta tesis muestra el resultado de haber estudiado la simplificación de textos centrada en las expresiones numéricas. Existen líneas de trabajo que no han sido tratadas y que se presentan como futuras líneas de continuación del trabajo presentado.

Nuestra metodología se centra en la simplificación de expresiones numéricas en los textos, sabiendo que hay otros muchos elementos dentro del texto que pueden ser objeto de simplificación. De ahí que hayamos presentado un

modelo genérico de simplificación de textos que nos permite decidir qué tipos de simplificaciones quieren realizarse. Como línea de trabajo futuro se plantea determinar otros tipos de simplificación, a nivel léxico o sintáctico, e instanciar el modelo presentado para llevar a cabo la simplificación de textos centrada en este tipo de simplificación. Para ello sería necesario implementar nuevas instancias del modelo con las herramientas necesarias y definir las reglas de simplificación basadas en hipótesis sobre el uso de técnicas de simplificación. Somos conscientes de que nuestro modelo depende de una variedad de factores, tales como el idioma del texto original, el tipo de texto, el usuario final para el que se está adaptando el texto y el nivel de dificultad deseado para el texto simplificado. Todos estos factores tienen que ser considerados para instanciar e implementar el modelo presentado.

Como mejora de las operaciones definidas e implementadas para la simplificación de expresiones numéricas, otra línea de trabajo futuro sería añadir representaciones gráficas de las expresiones numéricas. Estas representaciones ayudarían a comprender el significado matemático de la expresión numérica dada a través del uso de imágenes, gráficos o esquemas. Como una alternativa a la simplificación del texto, también consideramos la posibilidad de añadir información multimedia, como vídeo o audio, como una manera de ayudar al usuario final a leer y a comprender el texto original.

Otra línea de trabajo futuro incluye la evaluación de nuestras hipótesis para la representación de expresiones numéricas con experimentos reales con otros grupos objetivos aparte de los ya realizados para personas con dislexia. De esta forma mejoraríamos la personalización de las operaciones de simplificación que se pueden automatizar dependiendo del usuario final para el que se está simplificando. Además, una idea para cubrir de la mejor manera el modelado de usuario es implementar un sistema donde sea el propio usuario el que pueda configurar cada parámetro, y así personalizar, de manera individual, la tarea de simplificación automática de textos.

Los resultados obtenidos del caso de estudio real con personas con dislexia pueden llegar a ser de gran valor en la producción de las bases empíricas para el desarrollo o perfeccionamiento de las directrices para la simplificación de texto. Estas directrices existen en forma muy general (Freyhoff et al., 1998) y se emplean actualmente como referencia en una serie de esfuerzos para mejorar la accesibilidad de texto para grupos de usuarios con necesidades especiales. Una base empírica que relacione expresiones particulares con determinados grupos de usuarios sería una contribución muy positiva. Aunque la dislexia presenta manifestaciones heterogéneas entre los sujetos, encontramos patrones relacionados con la legibilidad y comprensión a partir de datos cuantitativos y cualitativos.

Otro campo en el que se puede esperar que estos resultados tengan un impacto es en el de la evaluación de la legibilidad. En términos generales, se utilizan modelos computacionales para predecir la legibilidad de los textos,

que se reducen a fórmulas matemáticas como Flesch, Flesch-Kincaid (Flesch, 1948) y SMOG (McLaughlin, 1969). Los esfuerzos actuales están considerando una serie de factores como el número promedio de caracteres por palabra y promedio de sílabas por palabra para predecir un resultado de legibilidad, pero no incluyen ninguna métrica específica para las expresiones numéricas. Con base en los resultados presentados aquí, podría hacerse un esfuerzo para ampliar el conjunto de funciones utilizadas en la evaluación de la legibilidad para incluir expresiones numéricas, ya que hemos visto que la presencia de información numérica influye en la lectura y comprensión del texto.

Como primera aproximación de trabajo futuro ya presentamos un caso especial de estudio de la representación numérica de los ingredientes en las recetas de cocina, ya que la representación matemática, las unidades y el lenguaje de la receta son factores que condicionan y transforman la información numérica. Los detalles del trabajo están en el trabajo de Bautista et al. (2013a).

Es posible lograr la accesibilidad universal si se toman en consideración los dispositivos asequibles, la tecnología, las cuestiones culturales y la falta de educación. Tenemos que seguir trabajando para lograr un diseño para la diversidad. En la diversidad es donde está la grandeza, y el diseño centrado en el usuario debe ser el principal objetivo de la accesibilidad universal.

Part II

Short version of the thesis in English: A Computational Model for Automatic Simplification of Numerical Expressions

This second part of the PhD Thesis is a condensed translation from Spanish into English of the previous PhD dissertation.

A Computational Model for Automatic Simplification of Numerical Expressions

Ph. D. Thesis

Susana Bautista Blasco

Departamento de Ingeniería del Software e Inteligencia Artificial
Facultad de Informática
Universidad Complutense de Madrid

Madrid 2015

A Computational Model for Automatic Simplification of Numerical Expressions

Report presented for the degree Ph.D in Computer Science

Susana Bautista Blasco

Supervised by

Prof. Dr. D. Pablo Gervás Gómez-Navarro

Prof. Dr. Raquel Hervás Ballesteros

Departamento de Ingeniería del Software e Inteligencia Artificial

Facultad de Informática

Universidad Complutense de Madrid

Madrid 2015

Chapter 8

Introduction

8.1. Introduction

We live in an "Information Technology Society", an expression that is becoming increasingly common, and is meant as a set of technologies, resources, procedures and techniques used in processing, access, storage and transmission of information in different formats. As a result of this society, there is a tendency to digitalize all kinds of information, such as recipes, payslips, news, etc, with the aim of making them more accessible to users. However, studies show that we are still far away from the ideal of a uniformly digitalized society where information is accessible to everyone.

The way in which information is written or presented can exclude many people whose level of reading skills makes them have problems in reading comprehension. There are several factors by which these skills can be affected, such as having had limited access to training, having social problems or having some cognitive disability. In addition, there are specific groups such as the deaf, the autistic, people with language disorders such as aphasia or dyslexia, people who are learning another language or the elderly, who have specific problems with reading. When submitting written information the diversity of the people who access it must be taken into account in order to make it accessible to all.

The Standard Rules on the Equalization of Opportunities for Persons with Disabilities from United Nations (UN, 1994) require governments to make all public information services and documentation to different groups of people with disabilities accessible to all people. The reason for this is because access to information for social and cultural development is a fundamental right to guarantee equality among people. One of the social problems that we are facing today is that there is difficulty in accessing information, as this information is presented in a way that makes reading and understanding the content of the information for different groups of less accessible society.

A first solution to this problem is the manual simplification of informa-

tion manually to adapt the difficulties of target users to whom it is directed. However, manual simplification is too slow and tedious to be efficient in producing the desired material. At the rate that advances in the information technology era are developing, where news travels through the network, distributed real-time through various means, it is not feasible to perform a manual simplification of information. Therefore, various attempts to automate part of this simplification process have been launched, focusing on the different transformations that can be applied in the process of text simplification.

Automatic text simplification is a relatively new task in Natural Language Processing. The aim of simplifying texts is to transform a text make it easier to understand for certain target users. In order to perform this task, mainly aimed at syntactic and lexical constructions that can be applied to the original text to generate a simplified version, researchers must identify what causes this difficulty in specific readers.

Users are directed to those texts which are obtained in the simplification process; these texts have very different characteristics, which diverge when an adaptation of the original texts is being performed. When we talk about content adaptation, we refer to the transformation of different contents that are difficult for the target user. The two main issues are what to adapt and how to adapt. The first question seeks to adapt elements in order to properly use the given content. To a great extent, the question of how to perform the adaptation depends on the characteristics of the users considered for adaptation. Content adaptation is performed in one way or another depending on the target user. Reading skills and the level of reading comprehension are affected by many external factors which influence the individual and social barriers such as poverty or lack of cultural training or access to advanced technologies. People with difficulties should read this proposal to solve a social problem, those left behind with the growth of digital information are getting increasingly older and in need of real-time solutions.

A case of information that creates difficulties for readers is numerical information. Many times, we access information that is represented in the form of numerical expressions such as economic, statistical, demographic data, numerical information on a recipe, a news article or a report. These numerical expressions can cause problems of understanding for many people for various reasons, either because they have disabilities or low academic training.

The survey of adult skills conducted in United Kingdom in 2011 as part of the Programme for the International Assessment of Adult Competencies (PIAAC) estimated that 7.5 million adults (only 22% of the population) are working at Level 2 or above in numeracy - roughly equivalent to a C on the GCSE maths examination for 16-year-old school children (Williams et al., 2003), (Miller y Lewis, 2012), (Williams et al., 2012). Roughly 2 in 5 people

(around 36%) said that poor maths skills had in some way held them back in their daily life. This rose to 4 in 5 for those who rated their numeracy skills as "poor" or "very poor". The other most common areas where people felt held back were in measuring and weighing (in cooking or administering medicine doses) and in understanding statistics in the media.

In Spain, the latest report from the Programme for the International Assessment of Adult Competencies (PIAAC) survey of adult skills[1], better known as the PISA report[2], evaluated performance in reading comprehension and math understanding among the population aged 16 to 65 years. It estimated that only 1 in 3 Spanish people are able to comprehend a long text or compare offers, and about 71.7% of adults can read and understand a simple text. In terms of numeracy, only 68.6% of adults are able to perform simple mathematical calculations and only 24.5% are able to interpret statistics, graphs or solve complex problems in steps. According to the study, the vast majority of Spanish people have difficulty extracting information from real mathematical situations like the comparison of tourism package deals, the calculation of the final price of a discounted purchase, and the interpretation of graphs and statistics.

With this goal, the main of this work is to carry out the automatic simplification of numerical expressions present in the text. The way in which information is presented can cause reading and comprehension problems for many people. Our work is based on the conclusions achieved by empirical study developed with experts. The adaptation of information is not an easy process but clearly necessary.

8.2. Motivation

In the area of automatic text simplification, we focus on a specific kind of information in order to adapt it and improve its readability and understandability. In our work, we have chosen numerical information because this kind of information causes problems for different groups of people in society.

As an example of this kind of problems we can see the daily news; this kind of texts presents all types of information. We can see that a lot of news has numerical information and the way in which information is presented affects the readability and comprehensibility of the text.

In our work we consider the expressions which represent quantities to be *numerical expressions*; optionally, they may have a numerical modifier, such as, for example *more than a quarter* or *almost 97%*, where *more than* and *almost* are the modifier in the expressions. This kind of expressions is found more frequently in information texts which present a lot of different

[1] https://www.mecd.gob.es/inee/Ultimosinformes/PIAAC.html

[2] http://www.mecd.gob.es/inee/estudios/piaac.html

numerical information.

For example, next we show a piece of news, taken from news agency Servimedia[3], and please note the amount and variety of numerical expressions used (highlighted):

> **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 país**. *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.*
>
> **ALMOST 400,000 PEOPLE** *IN PAKISTAN RETURN TO THEIR HOMES AFTER THE FLOODS*
>
> ***Around 390,000 people*** *have returned to their homes after they were forced to leave their houses due to floods caused by monsoon rains last summer in Pakistan. According to the UN Office for the Coordination of Humanitarian Affairs, this number accounts for* ***26%*** *of the* ***1.5 million Pakistanis*** *who had left their homes due to floods. On the other hand, the UN has managed to collect* ***34%*** *of the* ***2,000 million dollars*** *(****almost 1,400 million Euros****) asked for in an urgent call in the face of the catastrophe in Pakistan, the highest amount of money ever asked for by the UN as a result of a natural disaster. This catastrophe has killed* ***around 2,000 people****, has otherwise affected* ***more than 20 million of them****, and it has destroyed* ***close to 1.9 million homes*** *and* ***at least 160,000 square kilometres****, which is* ***a fifth of the country****. Given the set of circumstances, the UN Secretary General, Ban Ki-moon, has urged the international community to act "generously and swiftly" towards the humanitarian needs of Pakistan.*

[3] http://www.servimedia.es

In a relatively short text composed of five sentences, we find a total of 12 different numerical expressions, which is more than 2 expressions per sentence on average. These include expressions with quantities with or without modifiers, and so on. Such an information load, as well as the variety of different numerical expressions, may affect the reader's understanding of the text and prevent him from discovering cause and effect relations of the events presented in the news article.

8.3. Objectives

The access to available information for all is our main focus in this work, and in particular, the case of access to numerical information. In our approach we propose a generic model to carry out the automatic process of text simplification. From this model, we focus on the simplification of a special kind of information- numerical information. The objective is to make numerical information more accessible by rewriting difficult numerical expressions in a easier way. We propose a specific stage in the generic model to carry out this task. We need a set of rewriting strategies to achieve linguistically correct numerical expressions, easier to understand than the original and closer to the meaning of the original expression.

Next, we enumerate the main objectives of the present thesis:

1. Explore the text simplification area, focusing on a special kind of information: numerical information.

2. Present a specific model to simplify numerical expressions from the generic model to automatic text simplification.

3. Carry out an empirical study to identify the simplification strategies of numerical information.

4. Develop and implement different automatic simplification systems for numerical expressions for different languages following the model presented.

5. Evaluate the output of the systems developed.

8.4. Structure of the PhD

This dissertation is structured in seven chapters, the first of which is this introduction. Next we present the rest of chapters:

- **Chapter 9: Related Work.** In this chapter we present the main research related to this work, beginning with the task of simplifying

a text and the main tasks in text simplification. Next we present the main approaches to manual simplification and a review of automatic simplification approaches. Finally, we review generic natural language processing tools and specific tools for the treatment of numerical information.

- **Chapter 10: Theoretical bases for text simplification focused on numerical expressions.** This chapter shows the description and stages of the generic model for text simplification, and the specific model for numerical expression simplification. In addition, we present experimental identification of simplification strategies for numerical expressions in English and in Spanish carried out following different methodologies.
- **Chapter 11: Systems for the simplification of numerical expressions.** In this chapter we present the systems developed for the simplification of numerical expressions in English and in Spanish. In addition we show the evaluation carried out for each system.
- **Chapter 12: Discussion, Conclusions and Future Work.** In this chapter we discuss the work presented in this thesis. Finally, we present the main conclusions achieved through this research work and we show several lines of future work.

Abstract and Conclusions

In this chapter we have presented the concept of text simplification and the motivation of this thesis focuses on the simplification of numerical expressions to make information more accessible for people with special needs. In the next chapter we review the related work with this research.

Chapter 9

Related Work

In this chapter we present the task of text simplification and the main operations that is composed of. Next, a review of the manual approaches to text simplification is carried out and the automatic approaches developed in the area is summarized. We focus on the treatment of numerical information because it is the topic of this thesis. In addition, we review generic natural language processing tools and specific tools for the treatment of numerical information.

9.1. Text Simplification

The text simplification process was born of the need to adapt content texts for people who have difficulties reading and understanding a text in order to be functioning of society because access to information is a right for all persons. Text simplification consists of the transformation of a text into a similar text, but easier to read and understand. The objective is to achieve more accessible, attractive and communicative texts so that they are interesting and motive people with difficulties to read them. Access to reading is a social need and a recognized right and reading is a pleasure that lets people share ideas, thoughts and experiences.

30% of the population has reading difficulties which can be caused by different factors and this group of people needs a simplified version of texts in order access to the information. These factors may be intercultural difficulties, complex daily texts and cognitive aspects of the reader. People who may need a text adapted from the original version in order to understand its content are older people, people learning other languages, people with cognitive problems and a range of people with special educational needs (autism, aphasia, dyslexia, etc).

In order to communicate using written texts, it is important to use simple, clear and direct expressions in order to ensure better comprehension of the texts, to achieve good communication with the target user, to work

towards an inclusive social model. By carrying out certain operations at the lexical and syntactic level, linguistic complexity is lessened, thus obtaining a simplified text for the final user.

9.1.1. Main Tasks in Text Simplification

Text simplification includes main four tasks that have been researched over the years and the work done so far has covered these four objectives in one way or another. These four tasks are:

1. Syntactic simplification: transforming long and complex sentences into simple and independent sentences, splitting subordinate and main structures, changing sentences from passive to active, etc.

2. Lexical simplification: replacing complex vocabulary, considering context, using easier words or expressions, considering cases of polysemy (multiple meanings in a single word) and resolving ambiguity. Psycholinguistic data bases and thesauruses are often used.

3. Deletion of information: disregarding the information not needed to understand the main ideas in the text. Redundant information is deleted to help understand the text.

4. Clarification of information: adding explanations for concepts that are considered most difficult. Deciding what concepts are difficult but important and therefore should not be deleted, but should be defined or given the information necessary to help your understanding to help your understanding.

9.2. Manual Approaches to Text Simplification

There are several initiatives designed to develop the manual processing of text simplification following the European guidelines established by the IFLA (Freyhoff et al., 1998), published by Inclusion Europe Association (Inclusion Europe Association, 1998). All of these initiatives work in the area of Easy-to-read, a movement to create special material (books, documents, website, etc.), while tending the content and layout (format, margins, fonts, spacing, etc.) so that people with reading difficulties can read and understand the material.

These European guidelines are intended for authors, editors, information managers, translators and other people interested in generating information that is easy to read. Access to information is a fundamental aspect of participating in everyday life. Only informed individuals can influence or control the decisions that affect their lives. However, present structures deny access to information to a large number of people whose skills for reading, writing

or comprehension are diminished for different reasons. The purpose of the guidelines is to serve as a stimulus for generating readable documents so everyone in Europe can be integrated into the information society.

The general features of easy-to-read texts are:

- Using simple, direct language.
- Expressing one idea per sentence.
- Avoiding technical terms, jargon, abbreviations and acronyms.
- Structuring the text in a clear and consistent manner.

European guidelines cover some of the steps to follow for easy-to-read documents. There are two different options: making a text accessible, or generating a completely new text. In both cases, you have to start thinking about the target group and the main purpose that the text seeks to develop. With these two objectives in mind, we present the steps in the process of developing a readable text:

1. Define the purpose of the publication: what is meant and why it is important for people in the target group.
2. Address the issue of content: make a list of the key elements in the publication.
3. Draft the text: write the text based on the list of key issues.
4. Check that people understand the target group's first draft: before generating the final version of the document, reviewing it with real users helps to correct, improve and finish preparing the best possible version.

There are some general rules to be observed when writing a readable text:

- Use simple, direct language: use the simplest words expressed in the simplest way.
- Avoid abstract concepts: use concrete examples to facilitate the understanding of the topic.
- Use short words relating to everyday spoken language: avoid long words that are difficult to read or pronounce.
- Make the text as personal as possible: address readers directly and personally.

- Make use of practical examples that may be helpful for people to understand concepts and related information.
- Target readers respectfully: use adult language when writing for adults.
- Use short sentences mostly.
- Include one main idea per sentence.
- Use positive language: avoid denials and negative language, as it can cause confusion.
- Preferably use the active voice instead of the passive: Using active voice makes the document more lively and less complicated.
- Do not assume prior knowledge of the subject matter.
- Be systematic in using words; use the same word for the same thing.
- Choose simple punctuation marks: Avoid semicolons, hyphens and commas.
- Do not use the subjunctive: the uncertain future is vague and lends itself to confusion.
- Be aware if metaphorical language uses words that are not commonly used.
- Avoid of using numbers: long or complicated numbers are often incomprehensible. For small numbers, always use the number and not the word.
- Do not use words from another language.
- Avoid using references.
- Mention a contact address for further information, if possible.
- Avoid using jargon, abbreviations and acronyms. If it is unavoidable, always explain its meaning.

Next, we present the main approaches of manual text simplification. The initiative to adapt text to easy-to-read formats was born in Sweden in 1968 and is now based at the Fundation *Centrum för Lättläst*[1]. In 1984 they developed the first easy-to-read newspaper called *8Sidor* (8 pages). Its director, Bror Tronbacke, was the editor of the *Guidelines for easy reading material*, published in 1997 by the IFLA. The Swedish center is possibly the oldest and best organized in the world. Their experience has spread similarly

[1] http://www.lattlast.se/

Figure 9.1: European logo designed for easy reading *Inclusion Europe*

in neighboring countries, Norway and Finland. In Norway, the initiative is called *Leser søker bok* (reader looks for a book) and it was founded in 2003. In Finland there are two centers, one operating in Finnish and the other in Swedish, the country's co-official languages.

The organization Inclusion Europe[2] was founded in 1988; its headquartered is in Brussels and serves as the meeting point for the associations of people with intellectual disabilities in the European Union. It brings together organizations from 40 European countries and Israel. Its aim is to fight for equal rights and the full inclusion of people with intellectual disabilities and their families in all aspects of life. It designed a European logo (Figure 9.1) to identify all the texts that follow their guidelines. The organization writes and adapts readable texts in 20 European languages. An online magazine *e-Include*[3] featuring news, events and articles on different topics related to intellectual disability is published each day.

The Pathways I project[4] (2007- 2009) aims to formalize the need for easy reading as a tool for categorizing people with disabilities. Promoted by Inclusion Europe with partners from Austria, Germany, Finland, Ireland, Lithuania, Portugal and Scotland, it attempted to address the easy-to-read movement as a whole, not only considering the method of preparation and evaluation, but also thinking of people with intellectual disabilities as agents that compose texts. The idea continued with the Pathways II project[5] (2011-2013) expanding their materials to other European countries such as Croatia, Czech Republic, Estonia, Hungary , Italy, Slovenia, Slovakia and Spain.

The Easy Reading Association[6] with headquarters in Barcelona was the first of its kind to be established in Spain. It is a non-profit organization that works to bring reading to people with reading difficulties. It was created in 2002 and has more than 1500 subscribers, more than 122 easy-to-read books and 90 easy-to-read clubs to promote this activity among groups with reading difficulties.

[2] http://inclusion-europe.org/es
[3] www.e-include.eu
[4] http://inclusion-europe.org/en/projects/past-projects/pathways-i
[5] http://inclusion-europe.org/es/proyectos/pathways-ii
[6] http://lecturafacil.net

The website *Noticias Facil*[7] publishes readable articles, books and documents to bring the information to everyone. It is run by the ONCE Foundation[8] and targets people with intellectual or cognitive disabilities and people with reading comprehension problems. Keeping up with current events is very important, but there are people who do not understand the news published in newspapers because they have a complex style.

9.3. Automatic Approaches to Text Simplification

In all initiatives presented in the previous section the adaptation of texts was carried out manually. But simplifying a text manually is hard work in time and resources. Nowadays, information is generated very quickly and it is impossible to manually adapt accessible real-time texts. In order to solve this problem, automatic text simplification approaches have begun to appear.

In this section we present the main automatic simplification systems in chronological order, emphasizing their novelty and discussing how the field has progressed over time. In recent years, the idea has grown of applying machine translation in the process of simplifying texts, which is considered monolingual translation because it is a single language, but is translate from the original to a simplified version. Motivated by the new availability of simplified text corpora, there has been a dichotomy between systems designed manually with hand-written rules, and approaches based on corpora using statistical models. They have explored a variety of linguistic representations for the simplification of encoding operations, whether at the syntactical or lexical level.

One of the first important approaches was the work of Chandrasekar et al. (1996). Its motivation for text simplification was initially to reduce sentence length as a pre-processing step for a parser. Their second approach (Chandrasekar y Srinivas, 1997) was to have the program learn simplification rules from an aligned corpus of sentences and hand-simplified forms. The PSET (Practical Simplification of English Texts) (Carroll et al., 1998) project was perhaps the first to apply natural language technologies to create reading aids for people with language difficulties.

The other key foundational work in text simplification is the PhD dissertation of Dras (1999). He refers to the problem of reluctant paraphrasing, where text is altered to externally fit specified constraints such as length, readability or in-house style guides. The two key ideas here- synchronous grammars for monolingual paraphrasing and constraint satisfaction using integer programming- have been rediscovered in recent work on text simplification (De Belder et al., 2010), (Woodsend y Lapata, 2011), (Siddharthan

[7] http://www.noticiasfacil.es
[8] http://www.fundaciononce.es

y Angrosh, 2014).

Using rules based on simplification patterns, the system SYSTAR (SYntactic Simplification of Text for Aphasic Readers) was presented in Canning's writings (Canning, 2000). This system belongs to the PSET project. This module had to split sentences, change passive sentences to active sentences and resolve and replace anaphoric pronouns. For each sentence a recursive process of application was applied for each rule until all the rules had been applied.

Siddharthan's doctoral work (Siddharthan, 2003) focused on syntactic simplification. The most long-lasting contribution to the field was a detailed analysis of the discourse and coherence implications of syntactic simplification.

Inui et al. (2003) proposed a rule-based system to simplify English texts for deaf people. They defined rules at the syntactic and lexical level to apply them in the original text in order to generate a easier version for these people. Daelemans et al. (2004) applied automatic simplification at sentence level to generate subtitles in tv programs in Dutch and English for deaf people.

Williams y Reiter (2005) presented a text generation system that adapted its output for readers with low literacy. In the work of Petersen y Ostendorf (2007) an analysis of a parallel corpus of news was carried out in order to learn what kind of transformations people made for persons who are learning a second language.

The system *PorSimples* (Aluísio et al., 2008), (Candido et al., 2009) for Portuguese was developed in order to help low-literacy readers process documents on the web. With the development of the *Guidelines for materials in readable* IFLA (Freyhoff et al., 1998), in the work of Bautista et al. (2009) a subset of these guidelines was used to design and implement automatic rules at the syntactic and lexical level.

Zhu y Gurevych (2010) examined Wikipedia[9] in English and its simplified version Simple English Wikipedia[10]. Yatskar et al. (2010) focused on lexical simplifications carried out in the parallel versions and Biran et al. (2011) defined the complexity of a word as the proportion of its frequency in Wikipedia and in Simple English Wikipedia.

De Belder et al. (2010) used a rule-based system to simplify syntactic constructions such as apposition, relative clauses, subordination and coordination. Kandula et al. (2010) identified the difficult terms in the text and simplified them by using easier synonyms or adding an explanation of the term.

The main objective of the project *Simplext*[11] (Saggion et al., 2011) was to develop the product support for text simplification in Spanish for groups of

[9] http://en.wikipedia.org
[10] http://simple.wikipedia.org
[11] http://www.simplext.es/

people with special reading and comprehension needs. From a methodology of manual simplification defined by Anula (2007, 2008) it was possible to reduce the text complexity.

The problem of simplification was approached as a problem of automatic translation from English to English in the work of Coster y Kauchak (2011) following the methodology *Phrase Based Machine Translation (PBMT)*. Specia (2010) was the first to apply this methodology in text simplification for Portuguese. Wubben et al. (2012) improved the second step in PBMT with a different decoding stage.

Bautista et al. (2011c) presented an analysis of a parallel corpus to identify the transformation applied using the corpus created by Barzilay y Elhadad (2003). Walker et al. (2011) focused on lexical simplification and considered ambiguity as a factor to consider in the text simplification process.

Woodsend y Lapata (2011) presented a model based on quasi-synchronous grammar and integer linear programming. It used grammars to generate all possible rewrite operations for a source tree, and the integer linear programming to select the most appropriate simplification.

Bott et al. (2012) presented a system for Spanish text simplification. Aranzabe et al. (2012) proposed the first text simplification approach for Basque using specific rules for syntactic structures. In the case of French, we should point out the noteworthy work of Seretan (2012) in which they focused on reducing the syntactic complexity and the work of François y Fairon (2012) where they presented a new formula for measuring the readability of a text in French.

The *FIRST (Flexible Interactive Reading Support Tool)* project (Barbu et al., 2013) is developing a tool to assist people with autism spectrum disorders to adapt written documents into a format that is easier for them to read and understand. Saquete et al. (2013) developed a project focused on the treatment of educational texts in Spanish in order to reduce language barriers to reading comprehension for the hearing-impaired, or even people who are learning a language other than their mother tongue .

Recently, Siddharthan y Angrosh (2014) describe a synchronous dependency grammar for text simplification, that combines a manually constructed grammar for syntactic rules and an automatically acquired grammar for lexical rules and paraphrasing. In addition, the lastest approaches continue using handcrafted rules based on a typology of simplification rules extracted manually from a corpus of simplified French (Brouwers et al., 2014). Siddharthan (2014) reviewed discipline text simplification and highlighted most promising research directions to move the field forward.

We can see that in all systems of automatic simplification developed so far both the language with which they work, the target user, the kind of text and the level of difficulty to adapt the texts to play a key role in one way

or another. Each system considers a set of operations to simplify at various levels, syntactic or lexical, to carry out the adaptation of the original text. In following chapters we will see how these variables are considered in the work presented in this thesis.

9.3.1. Approaches Focused on Simplifying Numerical Information

Among the works on text simplification, we focus on those which deal with numerical information, since the work presented in this thesis is included within the simplification of numerical expressions. Here are the most important approaches in the research area of the processing of numerical information.

Bisantz et al. (2005) conducted a study to analyze the representation of probabilistic information. Research on linguistic probability (Budescu y Wallsten, 1995) have as a working hypothesis which in order to make decisions using linguistic representations, people people make these representations to numerical estimates with concrete values.

Peters et al. (2007) examined the concept of *numeracy*: why this is an important skill for providing health care, and what the best practices are for presenting numerical information in this context. To this end, they researched the influence of numerical information in understanding and what strategies exist to present numerical information to the patient.

The treatment of numerical information in the area of weather prediction was collected in the work of Dieckmann et al. (2009). They focused on decision markers that often occur with probability assessments. They conducted two studies to explore how decision markers vary in narrative and numerical information when making a prognosis.

Project *NumGen*[12] (*Generating Intelligent Descriptions of Numerical Quantities for People with Different Levels of Numeracy*) (Williams y Power, 2009, 2010) aimed to determine how to present the same numerical information in different ways for different users. To do this, they developed a system based on constraints in Prolog that given a ratio of input the system generates a set of possible equivalent versions in different mathematical representations. In addition, as part of the project, they constructed a corpus of newspaper articles that were high in numeric expressions.

Previous studies have shown that people choose accurate information instead of diffuse information, because it gives them a sense of security and makes their environment more predictable. The work of Mishra et al. (2011) showed that fuzzy environments with vague information (ranges) can help individuals make better comparisons of information that if the information is given precisely.

[12] http://mcs.open.ac.uk/sw6629/numgen/

In collaboration with the project *Simplext*[13] (Saggion et al., 2011) as part of this thesis, focusing on the treatment of numerical information, a component was developed based on rewrite rules for numeric expressions in the texts. A study was conducted to identify simplification strategies used to simplify numerical expressions from the parallel corpus, and a study was conducted with experts (Bautista y Saggion, 2014b).

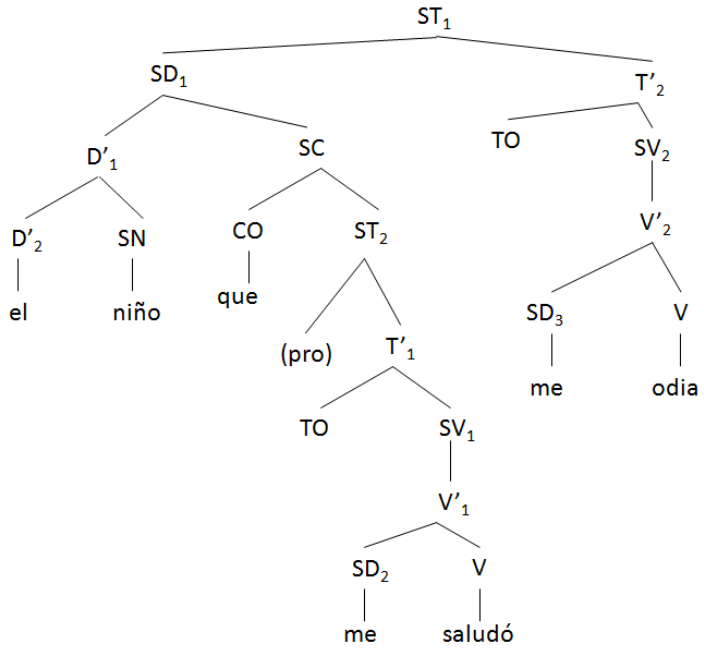


Figure 9.2: Example of a syntax tree for the sentence:*El niño que me saludó me odia*

9.4. Natural Language Processing Tools

Different tools and resources are important in the process of text simplification. In this chapter we present several tools of text analysis and we focus on the specific tools for the treatment of numerical expressions used in this thesis.

9.4.1. Syntactic Parsers

A natural language parser is a program that works with the grammatical structure of sentences. Known as statistical analyzers, they use the knowledge of the language, acquiring handmade analysis to try to produce the

[13] http://www.simplext.es/

most probable analysis of new sentences. These statistical analyzers still make some mistakes, but often work quite well. Its development was one of the greatest advances in natural language processing in the 1990s.

Within parsing, experts distinguish between constituent analysis and dependency analysis. Constituent analysis is characterized by the use of the inclusion relation (about phrases including others and, in the basic case they must construct phrases by lexical units). Given a sentence, this analysis constructs a syntax tree which is the representation of hierarchical relations among syntactic constituents. Figure 9.2 shows an example of a syntactic parse tree for a sentence. Dependency analysis is characterized by the use of binary relations (dependency) between lexical units. The words of a sentence depend on each other, so the direct object of a verb depends directly on it and an adjective depends on the noun. The purpose of this analysis is to build a dependency tree where each of the words in the utterance is represented and where the arcs between words represent the dependencies among them. Figure 9.3 shows an example of a dependency tree for a sentence. The use of one or the other depends on several factors, among which are the language with which you are working, the purpose and results of the work. Here are the main analyzers working in both English and Spanish.

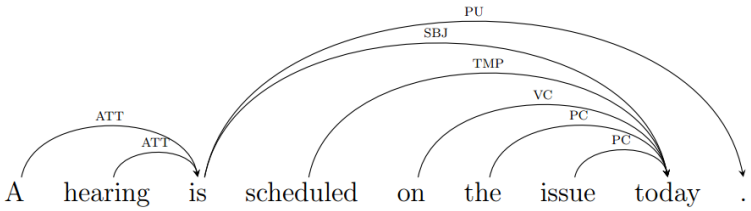


Figure 9.3: Example of a dependency tree for the sentence: *A hearing is scheduled on the issue today.*

For English one of the first dependency parsers is Minipar (Lin, 1998). Its coverage is quite broad. Another important parser for English is Stanford Parser (Klein y Manning, 2003) because it performs both dependency and constituent analysis. It was developed by Stanford University and it is implemented in Java. In addition to providing an analysis of English, the analyzer can and has been adapted to work with languages other than English. For example, it includes a Chinese analyzer based on the Chinese Treebank, the German-based analyzer Negra corpus and an Arab analyzer following Penn Arabic Treebank. It has also been used for other languages, such as Italian, Bulgarian and Portuguese. The analyzer provides an output dependency, as well as the structure of the constituent sentence trees.

Among the analyzers for Spanish the dependency parser JBeaver (Herrera

et al., 2007) stands out. It was developed using Maltparser (Nivre, 2003), a dependency analysis system. The most well-known and widely used Spanish analyzer is FreeLing (Padró et al., 2010) developed by the Polytechnical University of Catalonia at the TALP Research Center[14].

9.4.2. GATE

The tool GATE (General Architecture for Text Engineering) (Cunningham et al., 2002) is reused philosophy, not reinvented, so its main objectives are to integrate and interoperate with other systems and specific already-existing tools. It has a graphical interface and is integrated in a development environment that makes it easy to process different tasks and edit documents. GATE is free and its language-processing software uses specialized data structures and algorithms such as graphics annotation or finite state machines.

The set of integrated GATE resources is known as CREOLE (*a Collection of Reusable Objects for Language Engineering*). All resources can be exported as a *Java Archive* (.JAR) file plus an XML configuration file. When a set of resources has been developed, it may be included in a client application using *GATE Embedded*. GATE works with various document formats including XML, RTF, *email*, HTML, SGML and plain text. In all cases the format is analyzed and converted into a simple unified annotation model, generating a GATE document. GATE documents, corpora and annotations are stored in databases and can be visualized in the development environment.

GATE helps in the creation of these complex structures, the display of the processing results, and the measurement accuracy with regard to the results produced manually or semi-automatically. Figure 9.4 shows the GATE interface with different applications, language resources and processing resources to work on a text depending on the goals they have.

9.5. NLP Tools for the Treatment of Numerical Expressions

In order to simplify numerical expressions, specific tools to process this information for further processing are needed. In this section we present the different specific tools used in the thesis work. We reviewed two tools for English: a parser to analyze and label numerical expressions present in the text, and a program to approach proportions that allows us to obtain potential candidates for simplification given an input expression. For Spanish, we review the specific tool JAPE (*Java Annotation Patterns Engine*) created

[14] http://www.talp.upc.edu/

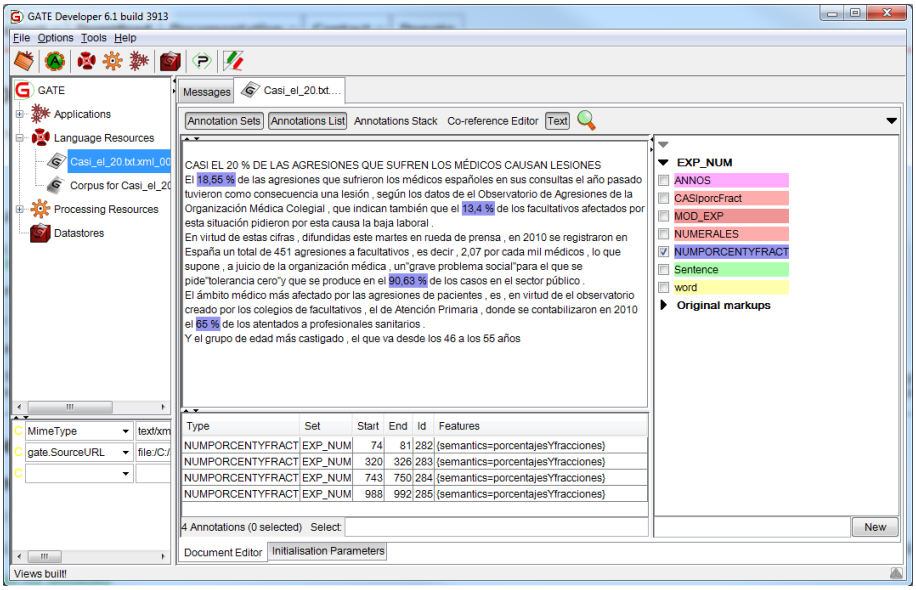


Figure 9.4: Example GATE interface for processing a text

by GATE and which has allowed us to define regular expressions to label numerical expressions in texts.

9.5.1. English Parser for Numerical Expressions

Sandra Williams developed a system that combines syntax and semantics to analyze and extract numerical expressions from texts in English (Williams, 2010). Thus, the system operates as a theoretical model of how numerical expressions are syntactically organized. The information extraction module performs semantic annotations in XML format in the expressions from the text.

The program is written in Java, and recognizes and labels numerical expressions in a text. It receives a plain text as input and generates a file in XML format as output with annotated sentences and numerical expressions using labels such as *<numex...>* and *</numex>*. It uses grammars with BNF format rules.

We show an example of the input text with numerical expressions marked in bold.

> *Maths and science comeback as A-Level grades soar*
> *A record number of students passed A-levels this year and more achieved A grades than ever before as the Government promised make the qualification tougher. The Joint Council for Qualifications published* ***827,737*** *grades for A-level this year, up from* ***805,657*** *in* ***2007****. A grades went to* ***25.9 per cent*** *of the en-*

*tries, up from **25.3 per cent** - and in Northern Ireland more than a third of students achieved an A. Girls continue to outshine boys at grades A-E, but the gap is beginning to narrow - down **0.3 per cent** at grade A. Entries for maths rose **7.5 per cent** from 2007, to **65,239**, while further maths was up **15.5 per cent**, to **9,483** entries. Less traditional subjects continued to increase in popularity with Chinese, Arabic and Russian showing steady increases every year since 2002. Some other languages suffered with a decrease in the number of students taking German, down **0.9 per cent** from 2007. But the number sitting French went up by **2.8 per cent** and there as a **1.5 per cent** rise in the number opting for Spanish. Sciences also fared well with entries for chemistry up **3.5 per cent**, physics up by **2.3 per cent** and biology up by **2.7 per cent**. Among the subjects showing increases were the sciences with entries for chemistry up **3.5%**, biology up **2.7%** and physics up **2.3%**. Dr Jim Sinclair, director, JCQ, said the record results were a cause for celebration. "These results are excellent and we congratulate all students on their achievement. The results show not only an improvement in the grades achieved but also an increased entry for mathematics, sciences and languages, which are positive and encouraging developments all round."*

Then we can see part of the XML output file generated where numerical expressions are annotated with labels generated by the analyzer.

```
<doctype w3c-doctype="numgen">
<header>
<title>
    Example XML markup for numerical expressions
</title>
</header>
<article id="1" topic="Maths and science comeback as A-Level
grades soar" date="14-Aug-08">

<sentence id="2">
    A record number of students passed A-levels this year
    and more achieved A grades than ever before as the
    Government promised make the qualification tougher.
</sentence>

<sentence id="3">
    The Joint Council for Qualifications published
    <numex id="1" type="ordinal" digits="yes" value="827,737">
        827,737
    </numex>
    grades for A-level this year, up from
```

```
    <numex id="2" type="ordinal" digits="yes" value="805,657">
        805,657
    </numex>
    in
    <numex id="3" type="date" digits="yes" units="year"
    value="2007">
        2007
    </numex>.
</sentence>

<sentence id="4">
    A grades went to
    <numex id="4" type="percentage" digits="yes" value="0.259">
        25.9 per cent
    </numex>
    of the entries, up from
    <numex id="4" type="percentage" digits="yes" value="0.253">
        25.3 per cent
    </numex>
    - and in Northern Ireland more than a third of
    students achieved an A.
</sentence>
...
</doctype>
```

Further details about this system can be found in Williams (2010).

9.5.2. *Proportion Approximation Program* in English

Given a proportion (value between 0 to 1), this program generates a set of equivalent versions. In addition for each version, it indicates the math type (fractions (F) or percentages (P)), the relation, the value expressed in the fraction and the kind of modifier that may be used together with the value in the percentage or fraction accordingly.

The development of this program was part of the *NumGen* project. It is a formal model for planning specifications for proportions (numbers between 0 and 1) and is formulated based on logical constraints. It uses generation grammars to express the various solutions generated from the input proportion in natural language.

Figure 9.5 shows an example of the output of the program for an input proportion. The detailed operation and design of the program are described in the work of Power y Williams (2012).

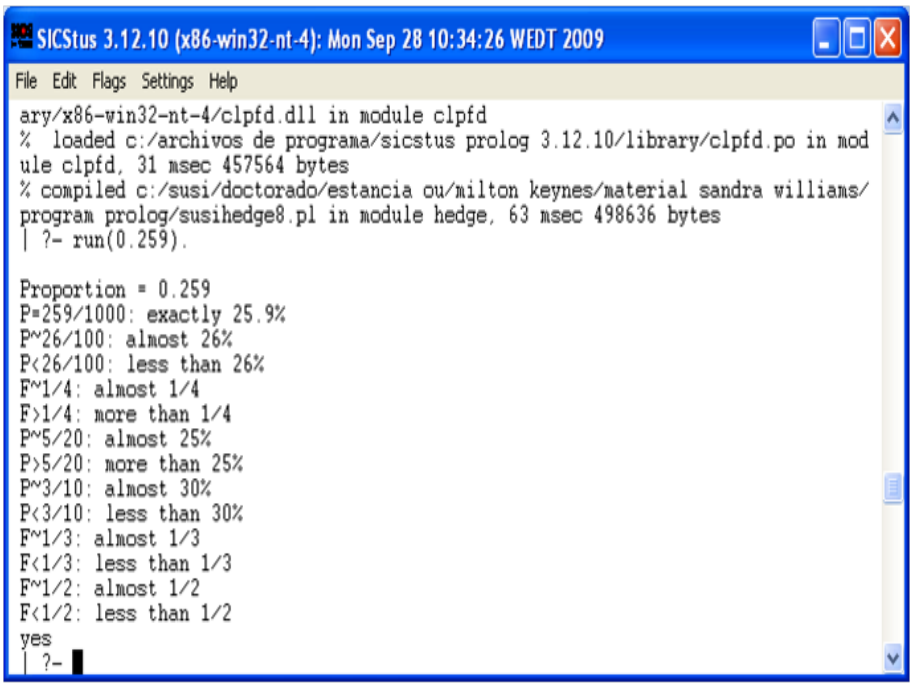


Figure 9.5: Example output of the proportion approximation program

9.5.3. JAPE (Java Annotation Patterns Engine)

JAPE (*Java Annotation Patterns Engine*)[15] belongs to GATE and recognizes regular expressions implemented in GATE annotated documents. JAPE is a version of CPSL- *Common Pattern Specification Language*[16].

JAPE grammars consist of a set of phases, each of which has a set of rules and patterns. These stages are executed sequentially and constitute a cascade of finite states in annotations. The left side of the rule (*Left-hand-side, LHS*) is formed by a pattern of annotation. The right side of the rule (*Right-hand-side, RHS*) consists of the manipulation statement annotation. The annotations on the left side can be referenced in the rules on the right side, using the tags defined in the pattern elements. The left part of the JAPE rule is relevant to what precedes the symbol "–> ", and the right part, what follows. When the left side matches the GATE score of a document, then the right side specifies what is to be done with the corresponding text.

[15] https://gate.ac.uk/sale/tao/splitch8.html#x12-2170008

[16] A good description of the original version of the language is in http://www.ai.sri.com/ appelt/TextPro/

Abstract and Conclusions

In this chapter we have presented the task of text simplification, focusing on numerical information. In addition, we have presented different natural language processing tools in order to use them in the process of text simplification.

In the next chapter, we present the theoretical bases for text simplification focused on numerical expressions. Also we present the generic model for text simplification and the experimental identification of the simplification strategies of numerical expressions carried out in order to decide what kind of transformations we have to implement in our systems for the automatic simplification of numerical expressions.

Chapter 10

Theoretical Bases for Text Simplification focused on Numerical Expressions

As we can see in the introduction chapter of this thesis, the task of text simplification for people with special needs it is really important. The manual simplifications proposed until now involve too much cost and effort. It is not a useful way to make the simplifications because there is plenty of volatile information nowadays.

In the text simplification process there are different kinds of transformations depending on what you want to simplify. In each stage, distinct information must be considered depending on the objectives and the purpose of the simplification. In our work we focus on the simplification of numerical expressions in text in order to help read and understand numerical information in the text.

One of the main objective in our work is the development of a computational model for automatic objectives simplification of numerical expressions and the variables to be considered in order to adapt the original text. In order to do so, we have studied, analyzed and decided what kind of transformations we need to implement. We carried out an experimental identification of the simplification strategies that people use when they have to adapt the numerical expressions in a text.

As we defined at the beginning of the work, we consider an expression representing a quantity to be a *numerical expression*: *53%* or *3489*, optionally modified by a numerical modifier such as *more than a quarter* or *around 97%* and sometimes accompanied by units such as *kms*, *liters* o *grams*. Furthermore, we consider different transformations that are used in the manual process of simplification to be simplification strategies. For example, changing the mathematical representation of the expression, using fractions rather than percentages, or using numerical modifiers when rounding the original

quantity, etc.

This chapter covers objectives 1, 2 and 3 from section 8.3. In section 10.1 the description and stages of the generic model for text simplification is addressed. In section 10.2 we present an instance of the generic model for simplification of numerical expressions. In section 10.3 the different methodologies proposed to carry out the empirical identification are presented. In sections 10.4, 10.5 and 10.6 various experimental identifications are addressed for each case.

The contents of this chapter correspond to the following publications: sections 10.1 and 10.2 to (Bautista et al., 2015), section 10.4 to (Bautista et al., 2011b) and (Bautista et al., 2011a), section 10.5 to (Bautista et al., 2012) and section 10.6 to (Rello et al., 2013). Some extra information, not always presented in the papers, is also referenced in this chapter.

10.1. Description and Stages of the Generic Model for Text Simplification

In this section we present a generic model of automatic text simplification, and describe its different working stages starting with the original text to be simplified. Figure 10.1 shows the stages of our generic model. We can see there are five variables that determine the configuration of the model in different stages: the language of the original text because it determines which tools can be used, the decomposition unit of the text, the type of text to be simplified, the target user that is performing the text simplification and the level of difficulty to which we want to adapt the final text.

Stage 1: Text Analysis: The input text is analyzed used NLP techniques, in order to find the information needed in the following stages.

Stage 2: Text Decomposition: From the previously analyzed text, the aim in this stage is to decompose the text into the linguistic units that are to be the target of the simplification process, such as words, sentences, or paragraphs.

Stage 3: Text Simplification: From the previous list of linguistic units, a set of simplification operations is applied to simplify them. There are different simplification tasks, such as syntactic transformations, where the structure of a sentence or a part of it is transformed; lexical substitutions, where only certain words are modified; the deleting of unnecessary information and the insertion of additional information.

Stage 4: Text Regeneration: In this stage, a recomposition of the text is addressed using the simplified versions of the target units in combination with the rest of the input text to reconstruct a whole, simplified version. This simplified text is the final output of the our model.

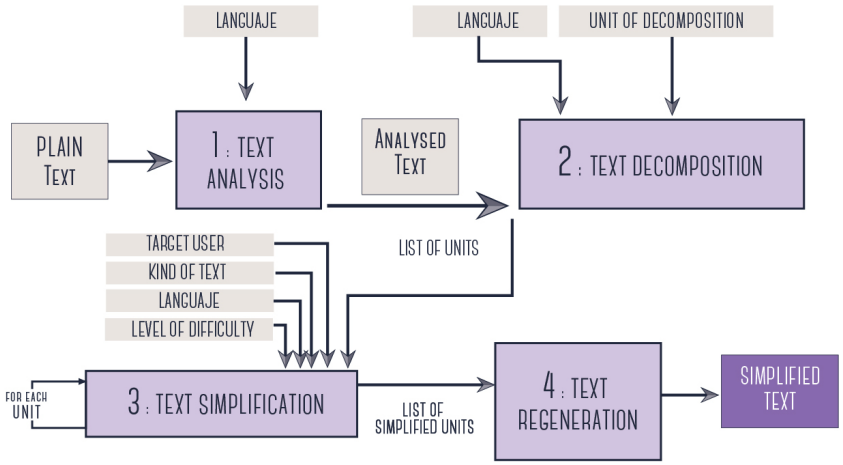


Figure 10.1: Stages of the Generic Model of Automatic Text Simplification. With plain text as input, the first stage consists of the analysis of the text. At the next stage, text decomposition is applied, which separates the original text into linguistic units. What follows is text simplification, which comprises different operations. Finally, text regeneration takes place and a simplified text is offered as system output.

In some cases, it may be necessary to combine more than one approach to simplification to achieve the desired result. When several simplification strategies have to be applied, an action arbitration has to be defined to decide the order in which they should be applied to the text. Combinations of radically different approaches- for instance, when summarization techniques based on the extraction of complete sentences are combined with lexical or syntactic simplification within the sentences may also require different instantiations of several stages.

More details about the generic model can be found in (Bautista et al., 2015).

10.2. Instance of the Generic Model for Simplification of Numerical Expressions

In each case you want to simplify a different type of information, it will require an instantiation of the generic model for text simplification discussed in the previous section. In our work we focus on the treatment of numerical information and for that reason the model is instantiated in a particular case to simplify numerical expressions in the texts. Furthermore, according to the objectives for which they work, the variables that are instantiated will have

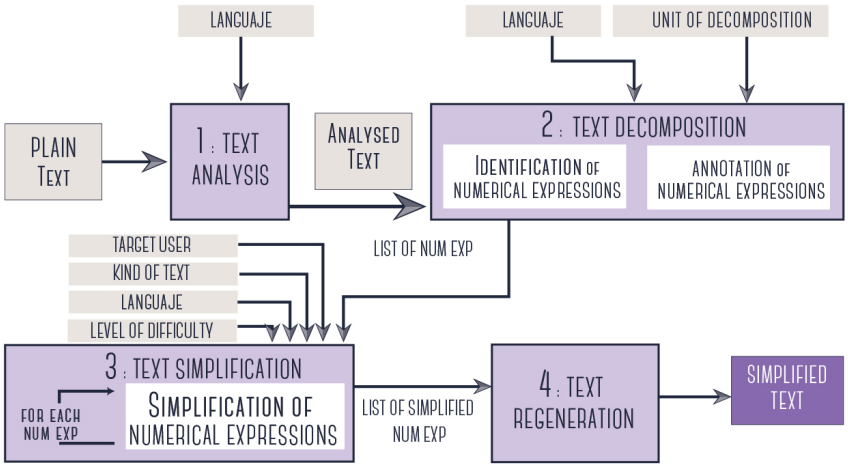


Figure 10.2: Stages of the specific model for simplification of numerical expressions

to be decided in the model for future computational implementation.

We pay special attention in stage 2 (Text Decomposition) of the model. In order to carry out the simplification of numerical expressions we split up this stage in two processes to identify and annotate the numerical expressions in the texts. Figure 10.2 shows the specific model.

Stage 3 (text simplification) focuses on simplifying numerical expressions. In this stage the simplification rules are defined and implemented. Finally, in text regeneration (stage 4) a final version with simplified numerical expressions is achieved. For this, the original numerical expressions are replaced by the simplified numerical expressions.

More details about the instance of the generic model for the simplification of numerical expressions can be found in (Bautista et al., 2015).

10.3. Methodologies Considered for Identifying the Simplification Strategies of Numerical Expressions

Pursuing the goal of obtaining a repertoire of simplification strategies for numerical expressions that can be implemented in an automatic simplification system, we carried out an experimental identification with experts in the field. When we refer to strategies, we consider the rules used by experts for the manual simplification process, then generalize this treatment and automate it in order to implement these rules in our simplification systems.

Next, we present the two parts in which the process of simplification of numerical expressions focuses, which are corresponding to the two parts that we identified in a numerical expression:

1. Use of modifiers: a quantity can be represented with a modifier or not, which determines the precision of the quantity. For example, *almost*, *more than* or *around*.

2. Quantity: expression that contains numerical information that is transmitted. For example, *24*, *98%*, *1/2*,

Optionally a quantity is written with metric units which may vary if the representation of the quantity changes. For example, *250 ml* or *1/4 l*. In this thesis we have not dealt with these units in the simplification process.

We propose a procedure for identifying different methodologies in each step of this process. Regardless of the language in which you work, the procedure is:

1. Propose some assumptions to be validated with experts.

2. Select the texts with which we will conduct the study.

3. Design a study where different design options can be considered.

4. Analyze the data collected.

Before going into detail about each of the steps in the proposed methodology, we will explain a number of concepts that are used in the assumptions:

- Common and uncommon values: the values are sorted according to frequency of use. There are many more known and common due to their high use (1/3, 50%, 1 in 4) than other, less common values (1/7, 69%, 1 in 34). This frequency of use makes the common values much more accessible to people with low numeracy training.

- Central and extreme values: the values are normalized in the range of 0.0 to 1.0, the central range (0.2-0.8) and the extreme ranges (from 0.0 to 0.2 and from 0.8 to 1.0).

- Modifier: quantifier accompanying numerical quantity to express its loss of precision.

- Error or loss of precision: the difference between the exact value of the quantity and the rounded value.

10.4. Experimental Identification with Experts of Simplification Strategies for Numerical Expressions in English

Following the procedure presented, in this section we show the experimental identification of simplification strategies for numerical expressions in English texts with the help of experts.

The range of numerical expressions is very large, so we mainly focus on three types of expressions: fractions, ratios and percentages. We consider using different mathematical representations for the quantity (fractions, ratios and percentages) with or without a modifier as simplification strategies.

10.4.1. Methodology for Numerical Expression Simplification in English

Our assumptions have been formulated as working hypotheses that we are going to validate by surveying experts.

In order to analyze the use of simplification strategies we define two working hypotheses. In regard to the use of modifiers and the loss of precision in the simplification process of numerical expressions, we define five more hypotheses. In all our hypotheses we consider the value of the original numerical expressions normalized between 0 and 1, and we refer to them as *proportions*.

Hypotheses related to simplification strategies:

H1: When experienced writers choose numerical expressions for readers with low numeracy, they tend to prefer round and common values to precise values. For example, *halves, thirds* and *quarters* are usually preferred to eightieths or forty-ninths, and expressions like *N in 10* or *N in 100* are chosen instead of *N in 365* or *N in 29*.

H2: The choice between different simplification strategies (fractions, ratios, percentages) is influenced by the value of the proportion, with values in the central range (say 0.2 to 0.8) and values at the extreme ranges (say 0.0-0.2 and 0.8-1.0) favoring different strategies.

Hypotheses related to the use of modifiers and loss of precision:

H3: The use of modifiers to accompany the simplified numerical expression is influenced by the simplification strategy selected. We consider the use of fractions, ratios and percentages as simplification strategies.

H4: The use of hedges to simplify the numerical expression is influenced by the value of the proportion, with values in the central range (say 0.2 to 0.8) and values at the extreme ranges (say 0.0-0.2 and 0.8-1.0) having a different use of hedges.

H5: The loss of precision allowed for the simplified numerical expression is influenced by the simplification strategy selected.

H6: There is some kind of correlation between the loss of precision and the use of modifiers, in such a way that the increase or decrease in the former influences changes in the latter.

H7: As a specific case of H6, when writers choose numerical expressions for readers with low numeracy, they tend not to use modifiers if they are not losing precision.

We carried out a study using a set of texts from the corpus of the project *NumGen* (Williams y Power, 2010). Our simplification process follows a scale of mathematical concepts defined from the learning levels in the *Mathematics Curriculum of the Qualifications and Curriculum Authority* (Department for Education, 1999). From this document, we have defined a scale of mathematical concepts to identify the levels of difficulty in understanding these concepts. This scale is the basis to define the difficulty levels considered in the simplification system (presented in section 11.1) for numerical expressions in English. Our survey took the form of a questionnaire in which participants were shown a sentence containing one or more numerical expressions which they were asked to simplify. The experiment was presented on SurveyMonkey[1], a commonly-used provider of web surveys. The survey was divided into three parts, simplification of numerical expressions for a person who can not understand percentages, for a person who can not understand decimals and free simplification of numerical expressions for a person with poor numeracy

More details about this experimental identification can be found in (Bautista et al., 2011b) and (Bautista et al., 2011a).

10.4.2. Data Analysis of the Simplification of Numerical Expressions in English

Since we divided our working hypotheses as to whether they were related to simplification strategies or the use of modifiers, we have carried out a data analysis for each subset collected in the survey.

10.4.2.1. Results of the Analysis of Simplification Strategies

In order to test hypothesis H1 (round or common values are preferred to precise ones), we carried out a series of two sample t-tests on common and uncommon fractions and ratios. The results support the hypothesis (no percentages: $p < .001$, no decimals: $p = .07$, free simplification: $p < .0001$, whole: $p < .0001$).

[1] www.surveymonkey.com

The use of different types of fractions seems to depend on the value being simplified, with quarters, thirds and halves (common fractions) preferred in the central range from 20% to 80%, and greater variety (and rarer use of fractions) in the periphery. This phenomenon can also be observed in non-numeric expressions. This was our hypothesis H2, and in order to test it we performed a series of two sample *t*-tests on the use of fractions, ratios, percentages and non-numericals in central and peripheral values. The results support the hypothesis (fractions: $p < .0001$, ratios: $p = .03$, percentages: $p < .0001$, non-numeric: $p < .0001$).

More details about this data analysis of simplification strategies can be found in (Bautista et al., 2011b).

10.4.2.2. Results of the Analysis of the Use of Modifiers

In order to test hypothesis H3 (the use of hedges in simplified numerical expressions is influenced by the simplification strategy selected), we carried out a series of two sample *t-tests* where statistical significance was adjusted for multiple comparisons by using the *Bonferroni correction*. The results do not support the hypothesis, as there is not a direct relation between the use of hedges and the strategy selected.

We performed another *t-test* adjusted by using the *Bonferroni correction* on the simplification strategies and central and peripheral values to test hypothesis H4 (the use of hedges to simplify the numerical expression is influenced by the value of the proportion, with values in the central range (say 0.2 to 0.8) and values at the extreme ranges (say 0.0-0.2 and 0.8-1.0) having a different use of hedges). The results show that the use of hedges is not influenced by central and peripheral values, rejecting our hypothesis H4 with a p-value $p=0.77$ in the worst case for the percentages strategy.

A new *t-test* adjusted by using the *Bonferroni correction* was done to test hypothesis H5 (the loss of precision allowed for the simplified numerical expression is influenced by the simplification strategy selected). The results seem not to support the hypothesis, as there is not a direct relation between the use of hedges and the loss of precision in the simplified numerical expression.

For hypothesis H6 (there is some kind of correlation between the loss of precision and the use of hedges), we looked for correlations between each part of the survey and each kind of simplification strategy. We carried out a non-parametric measure of statistical dependence between the two variables (loss of precision and use of hedges) calculated by *Spearman's rank correlation coefficient*. In general, the results show no correlation, so there is no linear dependence between the loss of precision in the strategy and use of hedges, rejecting our hypothesis.

Finally, when we analyzed hypothesis H7 (when writers choose numerical

expressions for readers with low numeracy, they tend not to use hedges if they are not losing precision), we worked with each part of the survey to study the cases where the loss of precision is zero and what the tendency of use of hedges is. With this data, it seems that we can accept hypothesis H7; that is, we found evidence for our assumption that when writers choose numerical expressions for readers with poor numeracy, they tend to use hedges when they round the original numerical expression, i.e when the loss of precision is not zero.

More details about this data analysis of use of hedges can be found in (Bautista et al., 2011a).

10.4.3. Summary of the Simplification Strategies for Numerical Expressions Identified in English

From the previous data reviewed and futher observed in the survey, we have identified the simplification strategies for numerical expressions in English. Thus, we present the main conclusions achieved:

1. Numerical expressions represented in words are transformed into their representation using digits.

2. Regardless of the level of difficulty, the common strategies identified are:

   - Percentages are rounded to the next value, for both common and uncommon values and for central and peripheral values.
   - No numeric expressions are used exclusively for the extreme values of the quantity.
   - Common fractions are used in the central range and other kinds of strategies are used in the extreme ranges.

3. Considering the level of difficulty for which it is simplifying:

   - If the level of difficulty corresponds to a person who does not understand percentages, the strategy to use is changing the expressions in percentages into their equivalent representation in fractions.
   - If the level of difficulty corresponds to a person who does not understand percentages with decimals, the strategy to use is rounding the quantity of the expressions in percentages with decimals to the next value without decimals.
   - If the level of difficulty corresponds to a person with difficulties with numerical expressions in a general way, the more used strategies are fractions, then ratios. By adapting the original expres-

sions with these kinds of representations, their difficulty is simplified.

4. With the data collected we have not observed a clear behavior in order to use ratios as a simplification strategy, or in relation to common or uncommon values or ranges in the central or peripheral proportion.

5. Regardless of the level of difficulty which simplifies, the use of modifier is not influenced by either the central or peripheral value of the proportion or by the simplification strategy used.

6. If there is no loss of precision in the simplification process, then a modifier is not used.

7. We have not found any correlation between loss of precision and the use of modifiers.

8. We have observed that if the original expression has a modifier, then the simplified expression kept the same modifier.

10.5. Experimental Identification with Experts of Simplification Strategies for Numerical Expressions in Spanish

For the case of identification of simplification strategies for numerical expressions in Spanish, different methodologies for study design have been explored in the general process proposed in this work. In all cases, the aim of the study is to draw conclusions about the kind of simplification operations that could be automatically applied to numerical expressions in Spanish.

In this case, three different methodologies are applied in order to design the study. In the first a parallel corpus of original and manually simplified texts is analyzed. In the second a survey with experts is carried out to identify the simplifications of numerical expressions preferred by them. In the third, we carry out a study with a specific group; in our case, people with dyslexia.

Among the myriad types of numerical expressions, we have limited our work to deal with monetary expressions (*15 millon of euros*), percentages (*24%*), fractions (*a quarter*), physical dimensions (*160,000 square kilometers*) and general quantities (*2,000 persons*).

The cases of the analysis of the parallel corpus and the survey with experts are both considered, because a comparative study is carried out and the results are discussed for both cases. The case with real users is presented as a separate case because the assumptions are different and the methodology applied differs slightly from the precious cases, hence it is contemplated as a separate case.

10.5.1. Methodology for Numerical Expressions Simplification in Spanish

We start with a set of assumptions that we raised in the process of simplification of numerical expressions in Spanish, to validate the results we get from the studies:

1. Original expressions expressed in letters must be rewritten by the corresponding digit version.

2. In the process of simplifying the numerical expression, if there is a loss of precision then a modifier is added and the original quantity of the expression is rounded.

3. If there is a modifier in the original expression but there is loss of precision in the simplification process, then the original modifier is changed and the quantity is rounded to generate the simplified version of the expression.

4. Numerical expressions are rewritten by changing their mathematical form; for example from percentages to fractions or from fractions to ratios.

Our study has two parts. One of them is the analysis of the parallel corpora of original and manual simplified texts; the other is the design,implementation and analysis of the survey with experts to extend our knowledge of possible simplifications of numerical information. On the one hand, we have the numerical expressions in context in the corpus where we can observe other kinds of simplification operation such as lexical transformations or syntactic changes. On the other hand, individual sentences were extracted from the corpus with numerical expressions and they were presented out of context to the participants of the survey to simplify them.

We have used the corpus from the project Simplext[2], a parallel corpus of 40 texts where we have identified the operations applied by text subjects in the simplification process. A subset of original texts was used to determine the sentences to be presented in the survey with experts. Different sentences were selected with different kinds of numerical expressions to increase the range of transformations applied by subjects when we ask them to simplify the numerical expressions. The survey was developed using the Google form-making tool, *Google Form* and we can access it in Google Docs[3].

More details about this identification and annotation of the numerical expressions in the corpus can be found in (Bautista et al., 2012).

[2] www.simplext.es

[3] https://docs.google.com/forms/d/1VG1G6voNbSVpP3gcGSHYMzou_-zX4xTFID8lG5hu984M/viewform

10.5.2. Data Analysis to Simplify Numerical Expressions in Spanish

We present a comparative analysis of the data collected and the validation of the assumptions with experts that was carried out.

The details of each analysis and the comparative analysis realized are presented in the article of Bautista et al. (2012).

10.5.2.1. Comparative Analysis of the Results

In order to make a comparative analysis of the results obtained in the study of the parallel corpus and the survey, we focus on the subset of numerical expressions used in the survey and the corpus. Later we extracted the set of operations applied in the simplification process and we compared the frequencies of use for this operations in the corpus and the survey.

In the results achieved from the corpus analysis more than 50% of numerical expressions were deleted while results from the survey suggested keeping the numerical information with a slight loss of precision by using rounded numbers with modifiers. In the survey, participants prefer rewriting the numerical information or keeping it the same way.

10.5.2.2. Validation of the Assumptions with Experts

We wanted to compare our assumptions (section 10.5.1) with the data collected in the study, both the results obtained from the analysis of the corpus and the responses to the survey by the experts.

For our first intuition (original expressions expressed in letters must be rewritten by the corresponding digit version), we can observe that this operation is considered in the parallel corpus but the participants in the survey do not consider this kind of transformation.

Second intuition (in the process of simplifying the numerical expression, if there is a loss of precision then a modifier is added and the original quantity of the expression is rounded), corresponds to the most applied operation in the survey (33.3 % usage). In contrast, in the case of the parallel corpus, adding a modifier as an applied transformation is not contemplated.

For our third intuition (if there is modifier in the original expression but there is loss of precision in the simplification process, then the original modifier is changed and the quantity is rounded to generate the simplified version of the expression), this transformation is contemplated in both cases, in the parallel corpus (with 7.4% of use) and in the survey (with 22.2% of use). Whether we are simplifying with or without context, this operation is important to implement in the process of simplification of numerical expressions.

Regarding the last intuition (numerical expressions are rewritten by chang-

ing their mathematical form, for example from percentages to fractions or from fractions to ratios), there are different cases of rewriting in the data collected and analyzed from the survey. These transformations have a frequency of use of 18.5%. However; in the corpus analysis there is not this kind of transformation in the process of simplification.

10.5.3. Summary of the Simplification Strategies for Numerical Expressions Identified in Spanish

We conducted a study from the parallel corpora of original texts and their corresponding simplified manual versions, along with a survey with experts that we asked to simplify numerical expressions present in sample sentences. From both empirical studies, simplification strategies for numerical expressions identified for the Spanish are:

1. In simplification with context, we observed:
    - Numerical information in parentheses is deleted.
    - In many cases, numerical information is removed directly rather than trying to simplify.
    - Expressions represented in letters were exchanged for expressions represented in digits.
    - Using context information, sometimes expressions are completely rewritten with no numerical expressions.
    - The quantity in the expression is rounded and sometimes a modifier is added or changed to remedy the loss of precision.
2. In simplification without context, we observed:
    - If the original expression has a modifier, it is usually changed and the number is rounded. Sometimes the same modifier is allowed.
    - There are very few cases of rewriting, since no context is accessed.
    - If the original expression has no modifier, a modifier is then added and the quantity is rounded.

In both cases we have observed that numerical expressions sometimes are not modified. This usually occurs in cases where the numerical expression is simpler itself, because it is a common value or a rounded value, such as 1/4 or 50%.

10.6. Experimental Identification with People with Dyslexia of Simplification Strategies of Numerical Expressions in Spanish

The other methodology considered in our work was to carry out the experimental identification of the simplification strategies of numerical expressions in Spanish with real users. In our case, we designed a study for a specific group of people, people with dyslexia. For this group we wanted to test:

- If rounded numbers are easier to read and understand than exact numbers.
- If readability and understandability were different when numbers were represented in fractions or percentages.
- If numbers represented in digits were more readable and understandable than numbers represented in words.

Before presenting each part of the study, let us define two terms we use for our study:

- *Readability*: attribute indicating that the representation of the information can be easily read.
- *Understandability*: capacity or ability to understand what is being read.

The aim of our study was to measure how numerical representation affects the readability and comprehension of a text for native Spanish speakers with and without dyslexia. We present the methodology used in the study. This work was carried out in collaboration with Dr. Luz Rello, from the research group TALN at the Universitat Pompeu Fabra in Barcelona.

More detail about this work can be found in (Rello et al., 2013).

10.6.1. Methodology for Numerical Expressions designed for People with Dyslexia

In order to carry out the study with people with dyslexia, we formulate our assumptions as a working hypothesis. Next we present the hypothesis for the experiments:

HD1.1: Readability will increase if digits instead of words, are used to represent numerical expressions. (*20* vs. *twenty*)

HD1.2: Understandability will increase if digits instead of words, are used to represent numerical expressions. (*20* vs. *twenty*)

HD2.1: Readability will increase if rounded numerical expressions, instead of unrounded expressions (with decimals), are used. (*48* vs. *48.3*)

HD2.2: Understandability will increase if rounded numerical expressions instead of unrounded expressions (with decimals), are used. (*48* vs. *48.3*)

HD3.1: Readability will increase if numerical expressions are expressed in percentages instead of fractions. (*25%* vs. *1/4*)

HD3.2: Understandability will increase if numerical expressions are expressed in percentages instead of fractions. (*25%* vs. *1/4*)

The material used in the study was a set of texts, created by Luz Rello and the author of this thesis, that considered features like length of the text, number of numerical expressions, number of named entities, etc. Each participant had to read several texts in Spanish with numerical expressions in different representations. We conducted three experiments with 72 persons (36 with dyslexia) using an eye-tracker and comprehension questionnaires to collect and analyze the data.

When reading a text, the eye does not move contiguously over the text, but alternates sarcades and visual fixations, i.e. jumps in short steps and rests on parts of the text. *Fixation duration* denotes how long the eye rests on a single place of the text. Fixation duration has been shown to be a valid indicator of readability. Shorter fixations are associated with better readability while longer fixations can indicate that processing loads are greater. Hence, we use fixation duration as measure to quantify readability. We use to measure text comprehension we used the questionnaires.

The details of the study, with the analysis and discussion of the results can be found in the (Rello et al., 2013).

10.6.2. Summary of the Simplification Strategies for Numerical Expressions Identified for People with Dyslexia

Once the study was conducted on people with dyslexia, from the working hypotheses raised and the analysis of data collected, we can say that regarding strategies to simplify numerical expressions, for this specific group (those with dyslexia), we have identified:

1. People with dyslexia read numerical expressions better in digits.

2. The simplification strategy of rounding the original quantity and adding a modifier to generate a simplified version of the numeric expression

increases reading time and does not improve understanding for people with dyslexia, as they have to read more.

3. Dyslexic people prefer to read numerical expressions represented in percentages, compared to representation in fractions, but find them more difficult to understand, because they have to infer the reference value in the percentage.

10.7. Comparison of Simplification Strategies for Numerical Expressions in English and Spanish

Experimental identification of simplification strategies for numerical expressions was carried out with different methodologies for English and Spanish. When these studies finished we identified a set of simplification strategies. Now we will present a comparison between the two languages studied in order to analyze what kind of strategies are common and what features there are in each language. These strategies there are the basis for the implementation of the simplification systems developed in this thesis and which we present in the next chapter.

Let us review the features of the simplification study for numerical expressions in English:

1. This study only considers the simplification of numerical expressions represented in percentages.
2. Non-numerical expressions are considered for the extreme ranges in the proportion.
3. Numerical expressions represented in fractions are simplified by using equivalent, more common fractions and adding a modifier.

Let us review the features of the simplification study for numerical expressions in Spanish:

1. This study considers a broader range with different kinds of numerical phrases: numerals, monetary expressions, percentages and fractions.
2. Non-numerical expressions are not considered as candidates in the simplified version.
3. Numerical expressions represented in fractions are not simplified by using equivalent fractions.

In both studies and regardless of the language to be simplified, we observed that:

1. Expressions represented in words must be transformed into their corresponding digit version.
2. Numerical expressions percentages are rounded to the nearest value to generate the simplified version of the expression.
3. Ratios are not changed in the process of simplification.
4. If there is no loss of precision in the simplification process, then modifiers are not used in the simplified version.
5. The use of modifiers in the simplified version of the expression includes several options:
   - If there is a modifier, then it is kept.
   - If there is no modifier, then one is added.

In addition, in the simplification study of numerical expressions for people with dyslexia in Spanish, we have identified that:

1. People with dyslexia read numerical expressions better in digits.
2. The simplification strategy or rounding the original quantity and adding a modifier to generate a simplified version of the numeric expression increases reading time and does not improve understanding for people with dyslexia, as they have to read more.
3. Dyslexic people prefer to read numerical expressions represented in percentages, compared to representation in fractions, but find them more difficult to understand, because they have to infer the reference value in the percentage.

Abstract and Conclusions

In this chapter we have presented the description and steps for the generic model for simplifying which we use in the implementation of the systems that will be presented in the next chapter.

In addition, we have presented different methodologies proposed for the identification of simplification strategies necessary to automate the process of simplification for numerical expressions. We have carried out different studies in English and Spanish to achieve the conclusions and some of them are used in the systems developed in this thesis.

In the next chapter we present the development and implementation of two systems to simplify numerical expressions in English and Spanish that follow the generic model presented and use the identified strategies in these two case studies conducted. For both systems an evaluation was performed

by experts that allowed us to evaluate the output of our systems and propose improvements therein.

Chapter 11

Systems for the Simplification of Numerical Expressions

In the previous chapter we set forth the theoretical basis for designing and implementing a process model and rules that allow us to automate the task of simplification, in our case focused on numerical expressions.

Our goal in this chapter is to present the implementation of two systems that validate the model presented to simplify numerical expressions, one for English texts and one for Spanish texts. For each system implemented, we will see the decisions made at each stage, the tools used and the variables set. Along with the description of each system, we present the evaluation that was carried out on it.

11.1. System for the Simplification of Numerical Expressions in English

In this particular system for English texts, the variable corresponding to the level of difficulty to which the numerical expressions must be adapted is considered at the stage where simplification operations are chosen. From the scale of learning mathematic concepts (presented in section 10.4.1), we considered three different difficulty levels in which the system can work: *Fractions Level* (simplification for people that only understand fractions), *Percentages without decimals Level (PWD)*(simplification for people that do not understand decimals) and *Percentages with decimals Level* (the most difficult level, where no adaptation is performed).

Using a graphical user interface (GUI) in Java, the user loads an original text, and chooses the level of difficulty; then a set of numerical expressions is automatically selected and a set of transformations is applied to adapt them, generating a text with the numerical expressions simplified at the chosen level as output of the system.

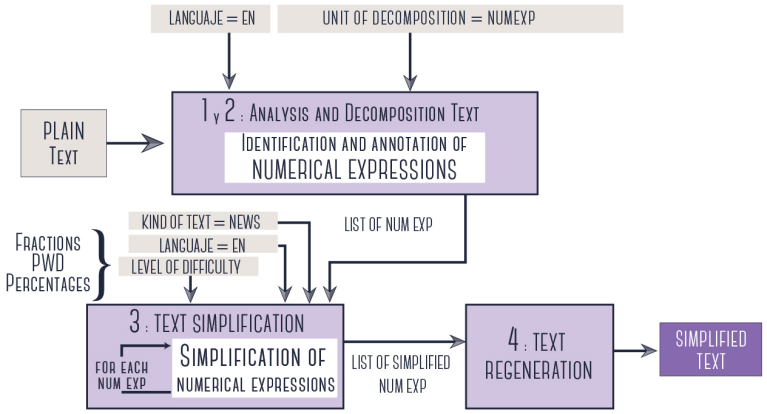


Figure 11.1: Stages of the automatic model for simplification focused on numerical expressions in English instanced for the simplification system for numerical expressions in English

The system only works with numerical expressions at the highest levels of difficulty; therefore, these kinds of expressions are percentages or decimal percentages in order to adapt them to easier levels. The user chooses the level of difficulty in the system interface.

In order to test how the system works, we used a text from the corpus of Project *NumGen* (Williams y Power, 2010). We decided to use the parser developed by Williams (2010) (presented in Chapter 9) that identifies and annotates the numerical expressions in a text. The system to simplify numerical expressions is implemented in Java, integrated with the *proportion approximation program* (Power y Williams, 2012) (presented in Chapter 9) that it is implemented in Prolog; we use it for the text simplification stage to calculate the candidates from among the numerical expressions to be simplified. The rules that we defined for what we learned from the experimental identification made are implemented in Java. System output is plain text with simplified numerical expressions. Figure 11.1 shows the specific stages of the model for simplification of numerical expressions.

More details about the system for simplification of numerical expressions at different levels of understandability can be found in (Bautista et al., 2013b).

11.1.1. Stage 1 and 2: Text Analysis and Text Decomposition

In our system stages 1 (Text Analysis) and 2 (Text Decomposition) of the model are performed simultaneously. In the analysis of the text the identification and annotation of the numerical expressions from the input text are

made. The text decomposition stage corresponds to the identification of the decomposition units to be simplified; in our case, the numerical expressions identified and annotated.

For the input text the system uses the parser developed by Williams (2010) to identify and annotate the numerical expressions in the text. The output of the parser is saved in an XML file and each numerical expression in the text is identified with the tag "*<numex>*". Different attributes are added to this tag to annotate the features of the identified expression, such as *type*, *format*, *given value*, *units*, *hedge*, *hedge-sem*.

11.1.2. Stage 3: Text Simplification

Let us remember the simplification strategies for numerical expressions in English (presented in section 10.4.3) identified in our study: we define and implement the rules for the system from them. The specific simplification rules for each case depend on the level of difficulty chosen by the user in the system gui. Thus, the system has to adapt each numerical expression identified and annotated in the previous stage to generate the simplified version. The simplification process has two stages: to obtain first the list of candidates and second the simplification rules where the modifier is chosen.

11.1.2.1. Obtaining the Candidate

In order to obtain the list of candidates we use a Prolog system, the *proportion approximation program* (presented in Chapter 9), which returns a list of candidates for substitution from the input value. This list is organized by the types of candidates, percentages or fractions, in decreasing order of precision with respect to the input value. The first option is the most precise for the type chosen. Depending on the level of difficulty chosen in the system, the first option which matches that level is chosen as a candidate to replace the original expression.

We can see an example of the process to obtain the candidate in Figure 11.2 at the level of difficulty *Fraction Level*, and the original numerical expression *more than 28%*. The original value is normalized by the parser and from the options of the *proportion approximation program*, the system chooses 3/10 as a candidate.

11.1.2.2. Applying Simplification Rules to Choose the Modifier

The use of the *proportion approximation program* ensures that the strategies selected in each case, percentages, fractions or non-numerical expressions, follow the ideas identified in the empirical study (section 10.4.3).

There is a strategy that it is always applied: numerical expressions represented in words are transformed to their representation using digits. The

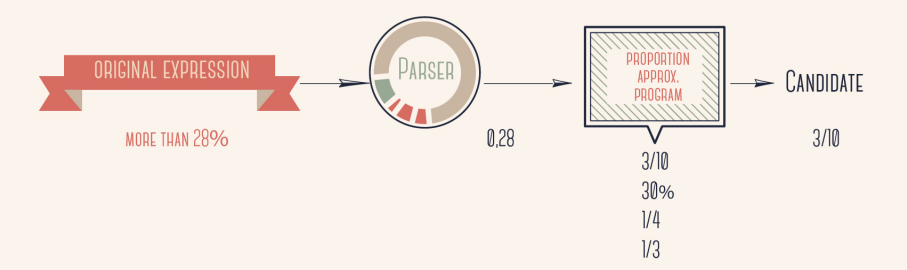


Figure 11.2: Obtaining eh candidate for simplification. The level chosen is *Fraction Level*, the original expression is annotated by the parser and this value is normalized. A candidate substitute value is chosen from the proportion approximation program.

system will apply a rule to carry out this transformation. In addition, we could see that there was not a clear use of ratios in the strategies identified so in our system ratios are not used as a candidate to be simplified from the original numerical expressions.

The system considers the level of difficulty chosen by the user then the rules corresponding are applied following the ideas identified in the study (section 10.4.3). From the set of strategies identified the rules used by the system carry out the simplification of the numerical expressions from the input text. The system has to adapt them to the level of difficulty chosen by the user. For each expression identified, the system only applies the simplification rules if the level of difficulty of the expression is higher than the level of difficulty chosen in the system. Then the system calculates the simplified version of the expression easiest to understand by the user. Following this idea and considering the conclusions obtained in the empirical study from chapter 10, we define two simplification rules focused on the kind of expression used:

- If the kind of numerical expression is *cardinal* or a *fraction*, and the format is *words* then the candidate to be used in the simplified version is the same number. For example, if the original expression is *six*, it will be replaced by *6*, or if the original expression is *a quarter*, the candidate is *1/4*.
- If the kind of numerical expression is *percentages* or *decimal percentages* and the format is *digit*, then the candidate is calculated by the *proportion approximation program*. In this case, the level of difficulty chosen in the system gui is lower than the level of difficulty of the numerical expression.

In order to finish the process of simplification, the system has to decide whether or not to use a modifier in the simplified version. A set of rules is

applied in each case and the main conclusions obtained in the empirical study are confirmed (section 10.4.3). The decision to use a modifier or not is based on the difference between the value of the original numerical expression and the value of the candidate.

As we did not find any correlation between the loss of precision and the use of modifiers, we decided that the system used modifiers if there was a loss of precision. We can see this decision in the set of rules defined in order to select a modifier in the simplified version.

More details about the use of modifiers can be found in (Bautista et al., 2013b).

11.1.3. Stage 4: Text Regeneration

This final stage of text regeneration is similar to the generic model, from the linguistically simplified units plus the rest of the original text, the system generates a simplified version. The output at this stage is a text where each numerical expression has been replaced by its simplified version.

The next text is the original text and following is the output of our system corresponding to the input text. The level of difficulty chosen in the system gui for the first text is *Percentages without decimals*, and for each numerical expression in the text the system calculates its simplified version. For the second text, the level of difficulty chosen is *Fractions* and for each numerical expression the system calculates its simplified version.

Original Text

Another record year for A-levels

The A-level pass rate rose for the 26th year in a row as record number of teenagers achieved top grades. But figures released by the exam boards highlighted startling discrepancies in Grade A pass rates between regions across England. Statistics from the exam boards showed greater improvements in students in the South East getting A grades in the past **six** years than those in the North East. The South East has seen a **6.1%** increase in A grades - to **29.1%** - since 2002 but the North East has seen an improvement of only **2.1%** - to **19.8%** - during the same period. But the percentage of pupils gaining passing E grades is rising quicker in the North East - an improvement of **3.4%** in **six** years compared with **2.8%** in the South East. Overall figures showed the national pass rate soared **above 97%** for the first time this year, while **one in four** **sixth**-formers were awarded A grades (**25.9%**, up from **25.3%** last year). The figures showed traditional subjects are still firm favorites with English and maths the top choices for candidates. Dr Mike Cresswell, director general of the AQA, said A-levels remained a "highly-valued qualification". He said he was particularly pleased to see the numbers of maths candidates rise from **60,093** last

year to **64,593** this year. "There was an upward trend that began a couple of years ago that has accelerated. There are more candidates doing mathematics than at any time in the past. It's important we have people with high mathematic skills so that has to be good news."

Simplified Text at *PWD* level

Another record year for A-levels

Last Updated: Thursday, 14 August 2008, 08:28 GMT The A-level pass rate rose for the 26th year in a row as record number of teenagers achieved top grades. But figures released by the exam boards highlighted startling discrepancies in Grade A pass rates between regions across England. Statistics from the exam boards showed greater improvements in students in the South East getting A grades in the past **6** years than those in the North East. The South East has seen a **around 6%** increase in A grades - to **around 29%** - since 2002 but the North East has seen an improvement of only **around 2%** - to **around 20%** - during the same period. But the percentage of pupils gaining passing E grades is rising quicker in the North East - an improvement of **around 3%** in **6** years compared with **around 3%** in the South East. Overall figures showed the national pass rate soared **above 97%** for the first time this year, while **1/4** **6th**-formers were awarded A grades (**around 26%**, up from **around 25%** last year) The figures showed traditional subjects are still firm favorites with English and maths the top choices for candidates. Dr Mike Cresswell, director general of the AQA, said A-levels remained a "highly-valued qualification". He said he was particularly pleased to see the numbers of maths candidates rise from **60,093** last year to **64,593** this year. "There was an upward trend that began a couple of years ago that has accelerated. There are more candidates doing mathematics than at any time in the past. It's important we have people with high mathematic skills so that has to be good news."

Simplified Text at *Fractions* level

Another record year for A-levels

Last Updated: Thursday, 14 August 2008, 08:28 GMT The A-level pass rate rose for the 26th year in a row as record number of teenagers achieved top grades. But figures released by the exam boards highlighted startling discrepancies in Grade A pass rates between regions across England. Statistics from the exam boards showed greater improvements in students in the South East getting A grades in the past **6** years than those in the North East. The South East has seen a **around 1/10** increase in A grades - to **around 3/10** - since 2002 but the North East has seen an improvement of only **around none** - to **around 1/5** - during the same period. But the percentage of pupils

> gaining passing E grades is rising quicker in the North East - an improvement of **around none** in **6** years compared with **around none** in the South East. Overall figures showed the national pass rate soared **around all** for the first time this year, while **1/4** **6**th-formers were awarded A grades (**around 1/4**, up from **around 1/4** last year) The figures showed traditional subjects are still firm favorites with English and maths the top choices for candidates. Dr Mike Cresswell, director general of the AQA, said A-levels remained a "highly-valued qualification". He said he was particularly pleased to see the numbers of maths candidates rise from **60,093** last year to **64,593** this year. "There was an upward trend that began a couple of years ago that has accelerated. There are more candidates doing mathematics than at any time in the past. It's important we have people with high mathematic skills so that has to be good news." Last Updated: Thursday, 14 August 2008, 11:28 GMT

11.2. Evaluation of the System for Simplification of Numerical Expressions in English

In order to know how our system works and evaluate the output, we carried out an evaluation of the system. We used a subset of sentences from the *NumGen* corpus. A questionnaire was presented to a set of human evaluators. The experiment was created and presented on SurveyMonkey[1], a commonly-used provider of web surveys.

For each original sentence, we presented two possible simplifications generated by the system. The first option generated by the system was for the *Fractions level*. The second option generated by the system was for *Percentages without decimals (PWD)*. Participants were asked to use their judgement to decide whether they agreed that the simplified sentences were acceptable for the original sentence. A Likert scale of four values was used to collect the answers.

We asked the same experts with which we performed the experimental identification to evaluate our system. We are grateful to all participants for their involvement in the experiment, first to identify the strategies and then to evaluate the automatic simplification performed by the system we implemented.

The answers from the participants were analyzed and evaluated. In total we collected 377 responses, 191 responses for the *Fraction Level* and 186 responses for *Percentages without decimals (PWD)*. Table 11.1 shows the use of central and extreme values and the use of common and uncommon values for each option presented in the survey. The values are the average

[1] http://www.surveymonkey.com/s/WJ69L86

Level	Total Average	Values	Average	Values	Average
Fraction	2,44	Central	2,87	Common	2,59
		Extreme	2,14	Uncommon	1,21
PWD	2,96	Central	3,00	Common	2,80
		Extreme	2,96	Uncommon	3,22

Table 11.1: System Evaluation: Fraction level and Percentages without decimals (PWD)

from the responses collected, using 1 to 4 for strongly disagree to strongly agree following the Likert scale presented in the questionnaire. For the two options presented in the survey, the average of the central values is higher than the average of the extreme values. As for the option of fractions, the average of the common values is clearly higher than the uncommon values. However, in percentages without decimals there is no significant difference between common and uncommon values.

In general, we can observe that the participants prefer common and central values for the fractions. However, there is no clear preference in percentages without decimals, because the important thing is deleting the decimals rounding the original quantity, regardless of whether it is central or extreme, common or uncommon. In addition, the experts think that the simplification done by the system in the *PWD level* is better than the simplification done in the *Fraction level*. They disagree specially with the simplification using fractions in two cases. One is the treatment of the extreme values where the system obtains "none" and "all" as possible candidates[2]. For example, the expression *1.3%* is simplified by *around none*. Another case is when uncommon fractions are used to simplify the numerical expression, like for example the expression *87.8 per cent* is simplified by *around 9/10*. In these two cases the average is lower than the rest of the averages achieved.

More details of the evaluation can be found in (Bautista et al., 2013b).

11.3. Simplification System for Numerical Expressions in Spanish

Following the generic model presented in section 10.1, we designed and implemented a system which applies the rules extracted from the empirical identification carried out to generate a Spanish text with simplified numerical expressions. Our system consisted of different components for each stage and

[2]See Power y Williams (2012) for a discussion of appropriate hedges for values near the extreme points of 0 and 1.

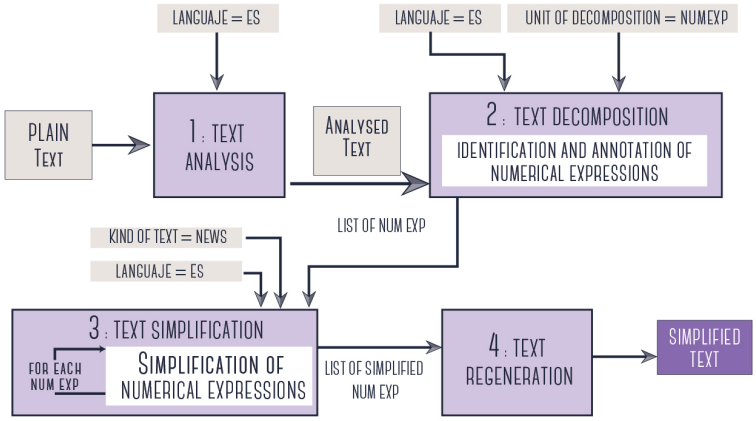


Figure 11.3: Stages of the automatic model of simplification focuses on numerical expressions in Spanish

they were integrated in a plug-in developed in Java and used in GATE. Figure 11.3 shows the specific stage in the model for Spanish text simplification. This system simplifies a broader numerical expressions such as numerals, monetary expressions, percentages and fractions.

More details about the simplification system can be found in (Bautista y Saggion, 2014a) and (Bautista y Saggion, 2014b).

11.3.1. Stage 1: Text Analysis

In this analysis stage two tasks are carried out: part-of-speech tagging and syntactic analysis of the Spanish text. The tool selected for both tasks is FreeLing (Padró et al., 2010).

Output of this stage is a list of analyzed sentences where each word has been labeled with morphologic information. This analysis is used in the next stage to carry out the identification and annotation of the numerical expressions in the text.

11.3.2. Stage 2: Text Decomposition

For each sentence the units have to be identified for the simplification process; in our case, these units are numerical expressions. In this stage, two different tasks are carried out, the identification of numerical expressions and their annotation.

In this stage different kinds of numerical expressions are identified in the text from the analysis carried out in the previous stage. FreeLing uses different labels for different kinds of numerical expressions following the annotation of EAGLES standard.

In our case, in order to annotate the different numerical expressions in the original texts, we used a set of JAPE grammars (*Java Annotation Patterns Engine*)[3]. JAPE is a version of CPSL - *Common Pattern Specification Language*. JAPE grammar is a set of rules, organized by phases and composed of patterns and actions. The output at this stage is a list of numerical expressions annotated with all the information needed for the simplification process.

11.3.3. Stage 3: Text Simplification

This stage receives a list of numerical expressions identified and annotated in the previous stage. The objective is to simplify them to generate a easier-to-read version of each numerical expression. In order to carry out this simplification a set of rules is designed and implemented from the simplification strategies identified in the survey for Spanish.

Our system considers the following strategies to be implemented:

1. Numerical expressions represented in words are changed for their representation in digits.

2. If the original numerical expression has a modifier, then it is kept in the simplified version and the quantity is rounded.

3. If the original numerical expression does not have a modifier, after applying a set of rules a modifier is chosen and added to the rounded quantity.

Simplification rules are implemented as follows: the quantity is always rounded and a set of rules is applied to select the modifier while considering the loss of precision. To obtain the rounded quantity, mathematical calculations are performed using different methods that form part of Java's *Math* package, which allows us to round the number to the nearest integer above original quantity. For example, if the original value of the amount is *0.891*, the system calculates the rounded value *1.0*. In order to choose the modifier for the simplified expression we define rules. If in the original expression already had a modifier, it is maintained and the quantity is rounded. For all other cases, the system compares the original quantity and the rounded quantity selected depending on the value or another modifier. The output is a list of simplified numerical expressions to use in the next stage.

More details about simplification rules can be found in (Bautista y Saggion, 2014b).

[3]https://gate.ac.uk/sale/tao/splitch8.html

11.3.4. Stage 4: Text Regeneration

The last stage is text composition, that is, from the simplified numerical expressions, with the rest of the text, a simplified version of the original text is generated.

The output of our system is a text where the numerical expressions have been simplified using a defined set of rules. Next, a post processing of the text is carrying out to solve some errors made during the treatment of the text by the parser FreeLing. We must also check the quotes, parentheses, slashes, hyphens and other punctuation marks that the analyzer makes, and introduce blanks.

Next we can see the original text and the simplified text generated by the system with the numerical expressions simplified.

Original Text

CASI EL 20% DE LAS AGRESIONES QUE SUFREN LOS MÉDICOS CAUSAN LESIONES

El **18,55%** de las agresiones que sufrieron los médicos españoles en sus consultas el año pasado tuvieron como consecuencia una lesión, seuán los datos de el Observatorio de Agresiones de la Organización Médica Colegial, que indican también que el **13,4%** de los facultativos afectados por esta situación pidieron por esta causa la baja laboral.

En virtud de estas cifras, difundidas este martes en rueda de prensa, en 2010 se registraron en España un total de **451** agresiones a facultativos, es decir, **2,07** por cada **mil** médicos, lo que supone, a juicio de la organización médica, un "grave problema social" para el que se pide "tolerancia cero" y que se produce en el **90,63%** de los casos en el sector público.

El ámbito médico más afectado por las agresiones de pacientes, es, en virtud del observatorio creado por los colegios de facultativos, el de Atención Primaria, donde se contabilizaron en 2010 el **65%** de los atentados a profesionales sanitarios.

Y el grupo de edad más castigado, el que va desde los **46** a los **55** años.

Simplified Text

CASI EL 20% DE LAS AGRESIONES QUE SUFREN LOS MÉDICOS CAUSAN LESIONES

El **casi 19%** de las agresiones que sufrieron los médicos españoles en sus consultas el año pasado tuvieron como consecuencia una lesión, según los datos del Observatorio de Agresiones de la Organización Médica Colegial, que indican también que el **más de 13%**de los facultativos afectados por esta situación pidieron por esta causa la baja laboral.

En virtud de estas cifras, difundidas este martes en rueda de prensa,

> en 2010 se registraron en España un total de **casi 500** agresiones a facultativos, es decir,**más de 2** por cada **1000** médicos, lo que supone, a juicio de la organización médica, un "grave problema social" para el que se pide "tolerancia cero" y que se produce en el **casi 91%**de los casos en el sector público.
> El ámbito médico más afectado por las agresiones de pacientes, es, en virtud del observatorio creado por los colegios de facultativos, el de Atención Primaria, donde se contabilizaron en 2010 el **más de 60%** de los atentados a profesionales sanitarios.
> Y el grupo de edad más castigado, el que va desde los **casi 50** a los **casi 60** años.

11.4. Evaluation of the simplification system for numerical expressions in Spanish

The simplification system for numerical expressions in Spanish was evaluated in two ways. The first was an intrinsic evaluation to analyze the linguistic accuracy of the system output. The second evaluation was conducted with experts to evaluate the system output directly.

11.4.1. Intrinsic Evaluation

In order to carry out this evaluation we used a subset of text from the corpus of the project *Simplext*. This subset has 57 texts, with 73 sentences.

The aim was to analyze the linguistic accuracy of the system output, verifying that the simplified sentence was correct and that the meaning of the sentence was maintained in the simplification process. For this, following evaluator's criteria (the author of this thesis) the original sentence and the simplified sentence were compared.

The results showed that out of the 73 sentences, in 61 cases the simplification of the numerical expressions was correct while in 12 sentences it vailed. 83.56% (almost 84%) of the simplified sentences were correct and preserved the meaning.

More details about this evaluation can be found in (Bautista y Saggion, 2014b).

11.4.2. Evaluation with experts

In order to carry out this evaluation we requested the participation of experts, primary and secondary teachers who work daily with students who need adaptations, and have academic training to evaluate the simplification made by our system. Our evaluation involved 42 experts, who enabled us to analyze the output of our system.

To perform the evaluation, we designed a questionnaire using the tool Google Form[4] which lets us create online forms and gather responses to questions. Participants were presented with 15 pairs of sentences, original and simplified by the system, with 34 numerical expressions of different types.

After analyzing the data collected in the questionnaire, the results showed that participants considered that the simplified version of the sentences generated by the system preserved the meaning compared to the original sentences with an average of 81.58% and a standard deviation of 9.24%. In addition, participants believe that sentences with simplified numerical expressions is grammatically correct with an average of 79.04% and a standard deviation of 12.98%. Finally, the evaluators considered numerical expressions were simplified correctly with an average of 72.69% and a standard deviation of 12.3%.

More details about the evaluation with experts can be found in (Bautista et al., 2015).

11.5. Comparison of the simplifying systems for numerical expressions implemented

Both systems carry out the simplification of numerical expressions in the text. Each has its own features and they share things in common. Therefore, we will make a comparison between the two systems.

The simplification system for English only simplifies numerical expressions in percentages. Although it is only one type of numerical expressions, the system offers the possibility of adapting such expressions at different levels of difficulty. In addition, because of the tools and resources used, the system uses no numerical expressions for the extreme values of the input proportion and is able to generate candidates fractions to simplify the original fractions in the text.

The simplification system for Spanish is for a broader range of numerical expressions such as numerals, monetary expressions, percentages and fractions, but does not identify different levels of difficulty in the process of simplification. Furthermore, it is not capable of generating non-numerical expressions to simplify certain expressions or inappropriate fractions with equivalent fractions in the process of simplification.

Both systems implement a rule to transform the expressions represented in words into their corresponding digit version. Numerical expressions represented in ratios are not treated in any of the two systems. Percentages are always rounded to the nearest value, correcting this loss of precision with the use of modifiers. Both systems have a set of rules to determine which modifier to use in the simplified version of the numeric expression being tested. In

[4] http://bit.ly/1wMwCwZ

general, if the original numerical expression has a modifier, then it is maintained. If there is no modifier and no loss of precision, then a modifier is added. If there is no loss of accuracy, no modifier is added.

As we can see, from the empirical study conducted following different methodologies the researches were able to define and implement simplification rules that allow the automatic simplification of numerical expressions in the texts to be performed. In this work we have presented two automatic simplification systems, one for English and one for Spanish. However, with our approach, it could carry out the task of simplification for any other language.

Abstract and Conclusions

In this chapter we have presented two systems to simplify numerical expressions, one for English and one for Spanish. Both systems have been implemented from the generic model specified in the previous chapter and whose rules were defined based on the strategies identified in the empirical study conducted. They have also been evaluated by experts who have examined the output generated by the systems for simplified numerical expressions.

Chapter 12

Discussion, Conclusions and Future Work

In this chapter we show the general discussion of the work presented in this thesis, the main conclusions and some lines of future work.

12.1. Discussion

Among the many options that are addressed in the field of text simplification, we had to decide to focus on a specific type, address the problem and propose a solution. Because until now numerical information from texts has hardly been treated in the area of simplification, the work of this thesis has focused on the processing of simplification of numerical expressions present in a text. From now on when the issue of simplification of numerical expressions is addressed, this study may provide the aspects identified that influence the automatic simplification process.

The work of this thesis presents a model for the automatic simplification of numerical expressions and the implementation of two computational systems that perform simplification for texts in English and Spanish.

The proposed generic model presented in chapter 10 covers a number of variables (kind of text, language, level of difficulty and target user) that are not always covered in the implemented systems. Each system presented in this thesis covers different aspects. For example, the English simplification system covers the level of difficulty in adapting a text type, the news, while the end user is a variable considered only for the case study in Spanish with people with dyslexia. We are aware that ideally we would have a system that covers all variables in the best way possible, i.e., all difficulty levels defined, the target user, the kind of text and language. This approach remains as future work for this researcher.

The tools and resources necessary to implement the generic model were

presented in chapter 9 and in particular the specific tools to simplify numerical expressions. The first decision to be made is the set of texts that will be used, hence the choice of the corpus is so important. In our work, both for English and for Spanish, we have the news corpus used in various research projects, which gave us the opportunity to use such material.

In order to define and computationally implement the rules to be applied in the automatic simplification of numerical expressions, an experimental identification with experts was carried out as we show in chapter 10. The procedure starts with the initial intuitions we want to validate with experts, then a selection is carried out of the necessary material, the study design is performed following different methodologies and finally the study is implemented; then the data collected is analyzed and our initial intuitions were validated or not.

The decision to use some tools or others at each stage of the model determines details of the study design and the system to be implemented (chapter 11). But once simplification strategies are identified, the process of defining and implementing them for the specific language is a process instantiation of the methodology identified and presented. These decisions are made depending on the language with which you are working and keeping in mind the purpose of the system as it is intended to be a helpful tool for people who have to adapt texts for people who have difficulty in reading and understanding the information they are accessing.

12.1.1. The Model as an Abstraction of Existing Practice

The generic model of text simplification presented in this thesis is intended as an abstraction that aims to cover a number of procedures being followed in practice by simplification systems already in existence. The language with which we are working determines the tools that are used in each stage to perform the process of simplification of numerical expressions. We pay special attention to the stage 2 model where the decomposition of the text is to make the identification and annotation of numerical expressions, and Stage 3 of simplification, which is implemented with rules run from lessons learned in the identification to determine experimental simplification strategies and the use of modifiers previously identified.

For example, the system presented by Carroll et al. (1998) in order to assist aphasic readers automatically simplifies English newspaper texts is available on the Internet. The system can roughly be divided into two main components: an analyzer component which provides a lexical tagger, a morphological analyzer and parser, and a simplifier component which subsequently adapts the output of the analyzer to aid readability for aphasic people using lexical and syntactical transformations. In terms of the model described in this paper, the analyzer component would correspond to Stage 1 of Text Analysis. Lexical transformation would correspond to a particular

instantiation of Stage 2 - Text Decomposition - to produce particular difficult words as target units, and a particular instantiation of Stage 3 - Text Simplification - that applies substitutions for these words to result in simpler alternatives for these difficult words. Syntactical transformations would correspond to a particular instantiation of Stage 2 - Text Decomposition - to produce particular syntactic constructions as target units, and a particular instantiation of Stage 3 - Text Simplification - that applies transformations to these syntactic constructions to result in simpler formulations. For both instantiations, a final process of reconstructing the complete version of the simplified text corresponds to Stage 4 - Text Regeneration.

Further systems can be analysed in a similar way. In the *Simplext* project Saggion et al. (2011) the text is analyzed using FreeLing Padró et al. (2010) and GATE Cunningham et al. (2002), which can be mapped onto Stage 1 of our model. Subsequent application of lexical and syntactical transformations can be considered as instantiations of Stages 2 and 3 as described above.

The *PorSimples* project Specia (2010) developed tools for Brazilian Portuguese and aims at developing technologies to make access to information easier for low-literacy individuals. This approach establishes that text simplification can be subdivided into syntactic simplification, lexical simplification, automatic summarization and other techniques. This proliferation of operations can be seen as the integration of several instantiations of our generic model, with different types of simplification operations being applied at different levels of granularity of decomposition (summarization at the level of the complete text, syntactic rewriting at the level of syntactic constructions, word substitution at the level of lexical terms). And the regeneration stage would be able to solve possible conflicts with different modules by proposing changes for the same text segment, applying some rules of priority or some kind of refereeing.

There are other types of systems which would not fit directly into the general model presented here. For example, some systems apply phrase based machine translations to the task of text simplification because they can only perform a small set of simplification operations such as lexical substitutions, deletion and simple paraphrasing. They are not well suited for reordering or splitting operations Coster y Kauchak (2011), Specia (2010). Other types are the systems which allow for some global optimization, such as integer linear programming. They use a synchronous grammar that combines a manually constructed grammar for syntactic rules and an automatically acquired grammar for lexical rules and paraphrasing Woodsend y Lapata (2011), De Belder et al. (2010), Brouwers et al. (2014), Siddharthan y Angrosh (2014). This kind of systems could be covered by our model with some modifications in the regeneration stage.

This analysis could be extended to other systems mentioned in section 9.3. We have shown how three different simplification systems for several

languages can be described in terms of our generic model of text simplification, and other specific systems would not fit directly into the general model. This can be taken as an indication of a certain degree of generality which may help to improve comparability across different systems. In each particular case, the language in which it will operate, the tools to be used, the kind of text and the target user have to be defined. Over these, each system applies its analysis and depending on the objective of the system, it defines its specific simplification transformations.

In addition, we can see that the simplification system architecture presented in the work of Siddharthan Siddharthan (2002) and the simplification generic model presented in this paper follow the same idea to generate the simplified version of an original text. Let us see the similarities and differences between the two proposals.

Siddharthan's work proposes an architecture consisting of three stages: analysis, transformation and regeneration. The first state provides the structural representation of a sentence and its part-of-speech tagging. The second stage uses transformation rules to generate plain text from the structure obtained by the previous state. And the third and final state is responsible for performing the syntactic simplifications referred to in each case.

Instead, the generic model presented in this work consists of four phases or stages: analysis, decomposition, simplification and text regeneration. Although our last phase is called the same as the third stage in Siddharthan's architecture, the functionality is not the same, since different operations are performed in one stage and the other. The first stage of our model is responsible for the text analysis at the syntactic level and part-of-speech tagging. The second stage decomposes the text, identifying the linguistic units that are to be simplified. The third stage is where simplification rules are applied to generate simplified versions of the units identified. And finally, the regeneration stage is responsible for reconstructing the text with simplified versions of the treated units to generate the final simplified text.

Comparing the two approaches, we can see that the initial state of analysis is common to both. The next state is different in both cases. In Siddharthan's architecture it is to generate plain text from the structures obtained in the analysis, while in the model proposed in this paper the second state corresponds to the identification of linguistic units to be simplified. The third state, where properly performed simplification transformations are performed, is called regeneration in Siddharthan's architecture while in this model it is called text simplification. The idea is the same in both cases, since they involve the simplification transformations according to certain rules. Furthermore, the model proposed in this paper provides one more state where the simplified text units are restored to generate the final simplified version of the text.

12.2. Conclusions

The changes in the *Information Technology Society* lead us to consider changes in the treatment and processing of information. For example, manual text simplification can not cope with the process of content adaptation for diverse audiences as it requires a lot of time and effort. This reality leads us to take advantage of technological solutions to help us to improve access to information for people with special difficulties.

Literacy, understood as written communication, involves cognitive reading processes that require effort and present difficulties for people with cognitive problems. The processing of numerical information plays a fundamental role in this literacy because numerical expressions are presented in different contexts, such as news, recipes, bills, etc. Our main motivation for automating the process of simplification of numerical expressions is the difficulty some people have to understand this information in a text.

In this work we have defined a generic model to carry out automatic text simplification, identifying the important variables that must be considered in the process. We focus on the treatment of numerical information as a special case study and validate our model by instantiating two real systems to simplify numerical expressions centered in text in English and Spanish. Both systems were evaluated by experts in the field, which have allowed us to analyze the collected results and to consider future improvements. Furthermore, in the case of Spanish, we present a real case study with people with dyslexia performed to test our hypothesis and which has allowed us to see up close the reality of a particular group, and to learn the specific strategies for this group's needs.

To achieve the simplification rules to be automated in each system, an experimental identification of simplification strategies used by humans was carried out. The identification process was conducted with experts, and thus we have information about the different transformations that can be applied and the use of modifiers when generating the simplified version of numerical expressions. Our studies show that the value of the proportion in the numerical expression influences on the strategy. The final mathematical form and the use of modifiers are important factors in the process of simplification of numerical expressions.

It is important to note that content adaptation is needed to cover the different levels that exist in the classroom educational level, so that information is accessible to more people. The purpose of the work presented is to help the experts to adapt content and be able to streamline this process so that the *Society of Information Technology* becomes a reality for all individuals who are part of it.

12.3. Future Work

The work described in this thesis shows the result of studying the simplification of texts focused on numerical expressions. There are lines of work that have not been treated and they are presented as future lines of work.

Our methodology focuses on simplifying numerical expressions in texts, while knowing that there are many other elements within the text may be simplified. Hence we have presented a generic simplification model for texts that allows us to decide what types of simplifications need to be done. As a future line of work we could extend the proposal to determine other types of simplification, at the lexical or syntactic level, and instantiate the model presented to perform the simplification of texts focused on this kind of simplification. This would require implementing new instances of the model with the necessary tools and defining rules for simplification based on assumptions about the use of simplification techniques. We are aware that our model depends on a variety of factors, such as the language of the original text, the kind of text, the target user that is adapting the text and the desired level of difficulty for the simplified text. All these factors have to be considered to instantiate and implement the model presented.

As an improvement of operations defined and implemented to simplify numerical expressions, another line of future work would be to add graphical representations of numerical expressions. These representations help to understand the mathematical meaning of the numerical expression given through the use of pictures, charts, or diagrams. As an alternative to simplifying the text, we also consider the possibility of adding multimedia information such as video or audio, as a way to help the end user to read and understand the original text.

Another line of future work includes evaluation of our hypothesis for the representation of numerical expressions with real experiments with other target groups beyond those already ddone for people with dyslexia. Thus, we would improve the personalization of simplification operations to be automated depending on the end user that is being simplified. In addition, an idea optimize user modeling is to implement a system where the user is the one who can set each parameter, and thus customize, individually, the task of the automatic simplification of texts.

The results of real case studies with people with dyslexia can be valuable in the production of the empirical basis for the development or refinement of guidelines for simplifying text. These guidelines are very general (Freyhoff et al., 1998) and currently used as a reference in a series of efforts to improve the accessibility of text for user groups with special needs. An empirical base that contained particular expressions relating to certain groups of users would be a very positive contribution. Although dyslexia presents heterogeneous manifestations among subjects, these are related to legibility

and comprehension from quantitative and qualitative data patterns.

Another field in which these results may have some impact is in the evaluation of readability. Overall, computational models are used to predict the readability of texts, which are reduced to mathematical formulas as Flesch, Flesch-Kincaid (Flesch, 1948) and (McLaughlin, 1969). Current efforts consider a number of factors such as the average number of characters per word and average number of syllables per word for readability to predict an outcome, but they do not include any specific metric for numerical expressions. Based on the results presented here, it could be an effort to expand the feature set used in the assessment of readability to include numerical expressions, as we have seen that the presence of numerical information influences the readability of the text.

As a first approach to future work, we are considering a special case study of the numerical representation of the ingredients in recipes, since the mathematical representation, units and language of recipes are factors that influence and transform numerical information. The details of the approach are at work Bautista et al. (2013a).

It is possible to achieve universal accessibility when you take into consideration affordable devices, technology, cultural issues and lack of education. We must continue working to achieve a design for diversity. Diversity is where greatness is, and user-centered design should be the primary goal of universal accessibility.

Parte III

Apéndices

Apéndice A

Publicaciones

En este Apéndice se muestran las publicaciones que han sido publicadas durante el desarrollo de la presente tesis.

A.1. Trabajos en simplificación de textos genérica

1. Bautista, S. y Gervás, P. Simplificación de texto para facilitar la comprensión lectora del usuario final. En *Conferencia de la Asociación Española para la Inteligencia Artificial (CAEPIA'09)*. Seville, Spain, 2009.

2. Bautista, S., Gervás, P. y Madrid, R. Feasibility analysis for semiautomatic conversion of text to improve readability. En *Proceedings of the Second International Conference on Information and Communication Technology and Accessibility (ICTA'09)*. 2009.

3. Ballesteros, M., Bautista, S. y Gervás, P. Text Simplification Using Dependency Parsing for Spanish. En *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval (KDIR'10)*, páginas 330–335. Valencia, Spain, 2010.

4. Bautista, S. y Gervás, P. Semiautomatic simplification to improve readability of texts for people with special needs. En *International Conference on Computers Helping People with Special Needs (ICCHP'10)*. Vienna, Austria, 2010.

5. Bautista, S., León, C., Hervás, R. y Gervás, P. Empirical identification of text simplification strategies for reading-impaired people. En *European Conference for the Advancement of Assistive Technology (AAATE'11)*. Maastricht, the Netherlands, 2011.

6. Bautista, S., Hervás, R. y Gervás, P. Accessible Numerical Information: Cookery Recipes as a Special Case. En *Proceedings of the*

Fourth International Conference on Information and Communication Technology and Accessibility (ICTA'13). 2013.

7. Hervás, R., Bautista, S., Rodríguez, M., de Salas, T., Vargas, A. y Gervás, P. Integration of lexical and syntactic simplification capabilities in a text editor. *Procedia-Computer Science Journal*, 2013.

A.2. Simplificación de textos centrada en expresiones numéricas en inglés

1. Bautista, S., Hervás, R., Gervás, P., Power, R. y Williams, S. How to Make Numerical Information Accessible. En *13th IFIP TC13 Conference on Human-Computer Interaction (INTERACT'11)*. 2011.

2. Bautista, S., Hervás, R., Gervás, P., Power, R. y Williams, S. Experimental identification of the use of hedges in the simplification of numerical expressions. En *Workshop on Speech and Language Processing for Assistive Technologies (SLPAT'11)*. 2011.

3. Bautista, S., Hervás, R., Gervás, P., Power, R. y Williams, S. A System for the Simplification of Numerical Expressions at Different Levels of Understandability. En *Workshop on Natural Language Processing for Improving Textual Accessibility (NLP4ITA'13)*. 2013.

A.3. Simplificación de textos centrada en expresiones numéricas en español

1. Bautista, S., Drndarevic, B., Hervás, R., Saggion, H. y Gervás, P. Análisis de la Simplificación de Expresiones Numéricas en Español mediante un estudio Empírico. *Linguamática*, vol. 4(2), 2012.

2. Drndarevic, B., Stajner, S., Bott, S., Bautista, S. y Saggion, H. Automatic Text Simplification in Spanish: A Comparative Evaluation of Complementing Modules. En *14th International Conference on Intelligent Text Processing and Computational Linguistics (Cicling'13)*. 2013.

3. Rello, L., Bautista, S., Baeza-Yates, R., Gervás, P., Hervás, R. y Saggion, H. One Half or 50 %? An Eye-Tracking Study of Number Representation Readability. En *14th IFIP TC13 Conference on Human-Computer Interaction (INTERACT'13)*. 2013.

4. Bautista, S. y Saggion, H. Can Numerical Expressions Be Simpler? Implementation and Demonstration of a Numerical Simplification Sys-

tem for Spanish. En *The 9th edition of the Language Resources and Evaluation Conference (LREC'14)*. 2014.

5. Bautista, S. y Saggion, H. Making numerical infomation more accesible: The implementation of a Numerical Expression Simplification System for Spanish. En *International Journal of Applied Linguistics 165:2 299-323 ISSN 0019-0829/ E-ISSN 1783-1490*. 2014.

6. Bautista, S. Hervás, R., Gervás, P. y Rojo, J. A Model for the Universal Access to Numerical Information. Aceptado en *Universal Access in the Information Society Journal ISSN 1615-5289/ E-ISSN 1615-5297*. En prensa. 2015.

Apéndice B

Estancias de investigación

Durante el período en el que se ha desarrollado la tesis doctoral se han realizado dos estancias de investigación:

1. La primera estancia de investigación se realizó bajo la supervisón de la Dr. Sandra Williams en el Computing and Communications Department, Centre for Research in Computing, The Open University, Milton Keynes, United Kingdom, del 8 de Junio de 2010 hasta el 30 de Septiembre de 2010. Durante este período se realizó la investigación de la simplificación de textos centrada en expresiones numéricas para el inglés.

2. La segunda estancia de investigación se realizó bajo la supervisión del Dr. Horacio Saggion, en el grupo TALN (Tractament Automatic del Llenguatge Natural) del Department of Information and Communication Technologies, de la Universitat Pompeu Fabra, Barcelona, España, del 1 de Marzo de 2012 al 1 de Julio de 2012. Durante esta estancia se realizó la investigación de la simplificación de expresiones numéricas en textos para el español.

En estas estancias se realizó parte del trabajo de investigación presentado en esta tesis. Además, fruto de las dos estancias, han surgido publicaciones y colaboraciones conjuntas.

Apéndice C

Charlas invitadas

Durante el desarrollo de la presente tesis, la doctoranda ha sido invitada a dar las siguientes charlas en las que ha presentado trabajos relacionados con su investigación:

1. "Semiautomatic Simplification to Improve Readability of Texts for People with Special Needs". En Flatlands Workshop. Oxford University. Annual meeting of the NLP groups at Cambridge, Essex, Open and Oxford universities. Junio 2010.

2. "Semiautomatic Simplification to Improve Readability of Texts for People with Special Needs". En Natural Language Generation Research Group. Computing and Communications Department, Centre for Research in Computing. The Open University, Milton Keynes, Reino Unido. Septiembre 2010

3. "Simplificación de Expresiones Numéricas en Español". En TALN Group (Tractament Automatic del Llenguatge Natural) at the Department of Information and Communication Technologies, Universitat Pompeu Fabra. Barcelona, España. Junio 2012.

Bibliografía

Aluísio, S. M., Specia, L., Pardo, T. A., Maziero, E. G. y Fortes, R. P. Towards Brazilian Portuguese Automatic Text Simplification Systems. En *Proceedings of the eighth ACM symposium on Document engineering*, DocEng '08, páginas 240–248. ACM, New York, NY, USA, 2008. ISBN 978-1-60558-081-4.

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

Anula, A. 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, 2008.

Aranzabe, M., Díaz de Ilarraza, A. y Gonzalez-Dios, I. First Approach to Automatic Text Simplification in Basque. En *Proceedings of the Workshop on Natural Language Processing for Improving Textual Accessibility (NLP4ITA) in LREC12*. 2012.

Ariles, C. y Jiménez, J. *Atención a la diversidad*. Consejería de Educacioón, Universidades, Cultura y Deportes. Dirección General de Ordenación, Innovación y Promoción Educativa. 2011.

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

Barbu, E., Martín-Valdivia, M. T. y Ureña-López, L. A. Open Book: a tool for helping ASD users' semantic comprehension. En *Proceedings of the Workshop on Natural Language Processing for Improving Textual Accessibility (NLP4ITA)*. 2013.

Barzilay, R. y Elhadad, N. Sentence Alignment for Monolingual Comparable Corpora. En *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, páginas 25–32. 2003.

BAUTISTA, S., DRNDAREVIC, B., HERVÁS, R., SAGGION, H. y GERVÁS, P. Análisis de la Simplificación de Expresiones Numéricas en Español mediante un estudio Empírico. *Linguamática*, vol. 4(2), 2012.

BAUTISTA, S., GERVÁS, P. y MADRID, R. Feasibility Analysis for Semi-Automatic Conversion of Text to Improve Readability. En *Proceedings of the Second International Conference on Information and Communication Technology and Accessibility (ICTA)*. 2009.

BAUTISTA, S., HERVÁS, R. y GERVÁS, P. Accessible Numerical Information: Cookery Recipes as a Special Case. En *Proceedings of the Fourth International Conference on Information and Communication Technology and Accessibility (ICTA)*. 2013a.

BAUTISTA, S., HERVÁS, R., GERVÁS, P., POWER, R. y WILLIAMS, S. Experimental Identification of the Use of Hedges in the Simplification of Numerical Expressions. En *Proceedings of the Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*. 2011a.

BAUTISTA, S., HERVÁS, R., GERVÁS, P., POWER, R. y WILLIAMS, S. How to Make Numerical Information Accessible. En *Proceedings of the 13th IFIP TC13 Conference on Human-Computer Interaction (INTERACT)*. 2011b.

BAUTISTA, S., HERVÁS, R., GERVÁS, P., POWER, R. y WILLIAMS, S. A System for the Simplification of Numerical Expressions at Different Levels of Understandability. En *Proceedings of the Workshop on Natural Language Processing for Improving Textual Accessibility (NLP4ITA)*. 2013b.

BAUTISTA, S., HERVÁS, R., GERVÁS, P. y ROJO, J. An Approach to Treat Numerical Information in the Text Simplification Process. *Universal Access in the Information Society, In press*, 2015.

BAUTISTA, S., LEÓN, C., HERVÁS, R. y GERVÁS, P. Empirical Identification of Text Simplification Strategies for Reading-Impaired People. En *Proceedings of the European Conference for the Advancement of Assistive Technology (AAATE)*. Maastricht, the Netherlands, 2011c.

BAUTISTA, S. y SAGGION, H. Can Numerical Expressions Be Simpler? Implementation and Demonstration of a Numerical Simplification System for Spanish. En *Proceedings of the 9th edition of the Language Resources and Evaluation Conference (LREC'14)*. Reykiavik, Iceland, 2014a.

BAUTISTA, S. y SAGGION, H. Making Numerical Information more Accessible: Implementation of a Numerical Expressions Simplification Component for Spanish. *ITL-International Journal of Applied Linguistics. Special Issue on Readability and Text Simplification. Peeters Publishers, Belgium*, vol. 165(2), páginas 299–323, 2014b.

BIRAN, O., BRODY, S. y ELHADAD, N. Putting it Simply: a Context-Aware Approach to Lexical Simplificaion. En *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. 2011.

BISANTZ, A. M., SCHINZING, S. y MUNCH, J. Displaying Uncertainty: Investigating the Effects of Display Format and Specificity. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 47(4), página 777, 2005.

BOTT, S. y SAGGION, H. An Unsupervised Alignment Algorithm for Text Simplification Corpus Construction. En *Proceedings of the Workshop on Monolingual Text-To-Text Generation*. Association for Computational Linguistics, 2011a.

BOTT, S. y SAGGION, H. Spanish Text Simplification: An Exploratory Study. *Procesamiento del Lenguaje Natural*, vol. 47, páginas 87–95, 2011b.

BOTT, S., SAGGION, H. y MILLE, S. A Text Simplification Tool for Spanish. En *Proceedings of the 7th International Conference on Language Resources and Evaluation. LREC'12*. 2012.

BROUWERS, L., BERNHARD, D., LIGOZAT, A. y FRANCOIS, T. Syntactic Sentence Simplification for French. En *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Targer Reader Populations (PITR) at EACL 2014*. Gothenburg, Sweden, 2014.

BUDESCU, D. y WALLSTEN, T. Processing linguistic probabilities: General principles and empirical evidence. *Journal Busemeyer. D.L.Medin and R.Hastie Eds. Decision making from a cognitive perspective. San Diego, CA. Academic Press*, páginas 275–318, 1995.

BUTTERWORTH, B. Foundational numerical capacities and the origins of dyscalculia. *Trends in Cognitive Sciences*, vol. 14(12), páginas 534–541, 2010.

CANDIDO, A., JR., MAZIERO, E., GASPERIN, C., PARDO, T. A. S., SPECIA, L. y ALUISIO, S. M. 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. Association for Computational Linguistics, Stroudsburg, PA, USA, 2009.

CANNING, Y. Cohesive Simplification of Newspaper Text for Aphasic Readers. En *3rd annual CLUK Doctoral Research Colloquium*. 2000.

CARROLL, J., MINNEN, G., CANNING, Y., DEVLIN, S. y TAIT, J. Practical Simplification of English Newspaper Text to Assist Aphasic Readers. En *Proceedings of the Workshop on Integrating Artificial Intelligence and Assistive Technology (AAAI)*, páginas 7–10. Madison, Wisconsin, 1998.

CHANDRASEKAR, R., DORAN, C. y SRINIVAS, B. Motivations and Methods for Text Simplification. En *Proceedings of the Sixteenth International Conference on Computational Linguistics (COLING '96)*, páginas 1041–1044. 1996.

CHANDRASEKAR, R. y SRINIVAS, B. Automatic Induction of Rules for Text Simplification. *Knowledge-Based Systems*, vol. 10, 1997.

CHUNG, M. H.-J. K. J., JIN-WOO. y PARK, J. Enhancing Readability of Web Documents by Text Augmentation for Deaf People. En *Proceedings of the International Conference on Web Intelligence, Semantics, and Mining (WIMS)*. 2013.

CLEMENTE, M. y DOMÍNGUEZ, A. *La Enseñanza de la Lectura: Enfoque Psicolingüístico y Sociocultural*. Colección Psicología/Ediciones Pirámide Series. Ediciones Pirámide, 2003.

COHEN, L., DEHAENE, S. y VERSTICHEL, P. Number words and number non-words: A case of deep dyslexia extending to arabic numerals. *Brain*, vol. 117, páginas 267–279, 1994.

COSTER, W. y KAUCHAK, D. Learning to Simplify Sentences Using Wikipedia. En *Proceedings of Text-To-Text Generation, ACL Workshop*. 2011.

CUNNINGHAM, H., MAYNARD, D., BONTCHEVA, K. y TABLAN, V. GATE: A framework and graphical development environment for robust NLP tools and applications. En *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*. 2002.

DAELEMANS, W., HÖTHKER, A. y SANG, E. T. K. Automatic Sentence Simplification for Subtitling in Dutch and English. En *Proceedings of the 4th International Conference on Languaje Resources and Evaluation*, páginas 1045–1048. 2004.

DE BELDER, J., DESCHACHT, K. y MOENS, M.-F. Lexical simplification. En *Proceedings of Itec2010 : 1st International Conference on Interdisciplinary Research on Technology, Education and Communication*. 2010.

DEPARTMENT FOR EDUCATION. Mathematics: the National Curriculum for England. Qualification and Curriculum Authority. Informe técnico, 1999.

DIECKMANN, N., SLOVIC, P. y PETERS, E. The Use of Narrative Evidence and Explicit Likelihood by Decision makers Varying in Numeracy. *Risk Analysis*, vol. 29(10), 2009.

DRAS, M. *Tree Adjoining Grammar and the Reluctant Paraphrasing of Text*. Tesis Doctoral, Macquarie University, Australia, 1999.

DRNDAREVIC, B., STAJNER, S., BOTT, S., BAUTISTA, S. y SAGGION, H. Automatic Text Simplification in Spanish: A Comparative Evaluation of Complementing Modules. En *Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics (Cicling)*. 2013.

DUBOIS, B. Something of the order of around forty to forty-four. *Language in Society*, vol. 16(4), páginas 527–541, 1987.

ELHADAD, N. Comprehending Technical Texts: Predicting and Defining Unfamiliar Terms. En *Proceedings of the AMIA Annual Symposium*. Washington, DC, 2006.

EVANS, R., ORASAN, C. y DORNESCU, I. An evaluation of syntactic simplification rules for people with autism. En *Proceedings of the Third Workshop on Predicting and Improving Text Readability for target reader populations*. 2014.

FLESCH, R. A new readability yardstick. *Journal of Applied Psychology*, vol. 32, páginas 221–233, 1948.

FRANÇOIS, T. y FAIRON, C. AI readability formula for French as a foreign language. En *Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing (EMNLP 2012)*. 2012.

FREYHOFF, G., HESS, G., KERR, L., MENZEL, E., TRONBACKE, B. y VEKEN, K. V. European Guidelines for the Production of Easy-to-Read Information for People with Learning Disability. Informe Técnico ISLMH, 1998.

HERRERA, J., GERVÁS, P., MORIANO, P., MORENO, A. y ROMERO, L. JBeaver: un Analizador de Dependencias para el Español Basado en Aprendizaje. En *Proceedings of the 12th Conference of the Spanish Society for Artificial Intelligence (CAEPIA 07), Salamanca, Spain*, páginas 211–220. Asociación Española para la Inteligencia Artificial, 2007.

INCLUSION EUROPE ASSOCIATION. Inclusion europe. *http://www.inclusion-europe.org*, 1998.

INUI, K., FUJITA, A., TAKAHASHI, T., IIDA, R. y IWAKURA, T. Text Simplification for Reading Assistance: A Project Note. En *Proceedings of the 2nd International Workshop on Paraphrasing: Paraphrase Acquisition and Applications*, páginas 9–16. 2003.

JUST, M. y CARPENTER, P. A theory of reading: From eye fixations to comprehension. *Psychological review*, vol. 87, páginas 329–354, 1980.

KANDULA, S., CURTIS, D. y ZENG-TREITLER, Q. A Semantic and Syntactic Text Simplification Tool for Health Content. En *Proceedings of the AMIA Annual Symposium*. 2010.

KLEIN, D. y MANNING, C. D. Fast Exact Inference with a Factored Model for Natural Language Parsing. En *Advances in Neural Information Processing Systems 15*, páginas 3–10. MIT Press, 2003.

KRIFKA, M. Be brief and vague! And how bidirectional optimality theory allows for Verbosity and Precision. En *Sounds and Systems: Studies in Structure and Change: A Festschrift for Theo Vennemann (Trends in Linguistics 141)*, páginas 439–458. Mouton de Gruyter, Berlin, 2002.

LANDERL, K., BEVAN, A., BUTTERWORTH, B. ET AL. Developmental dyscalculia and basic numerical capacities: A study of 8–9-year-old students. *Cognition*, vol. 93(2), páginas 99–125, 2004.

LIN, D. Dependency-based evaluation of MINIPAR. En *Proceedings of Workshop on the Evaluation of Parsing Systems*. 1998.

LOZANOVA, S., STOYANOVA, I., LESEVA, S., KOEVA, S. y SAVTCHEV, B. Text Modification for Bulgarian Sign Language Users. En *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*. Association for Computational Linguistics, Sofia, Bulgaria, 2013.

MACKAY, D. *Sustainable Energy - without the hot air*. UIT Cambridge Ltd, 2009.

MCCLOSKEY, M., CARAMAZZA, A. y BASILI, A. Cognitive mechanisms in number processing and calculation: Evidence from dyscalculia. *Brian and Cognition*, vol. 4, páginas 171–196, 1985.

MCLAUGHLIN, G. H. SMOG Grading - a New Readability Formula. *Journal of Reading*, vol. 12(8), páginas 639–646, 1969.

MILLER, G. A., BECKWITH, R., FELLBAUM, C., GROSS, D. y MILLER, K. J. Introduction to WordNet: An On-line Lexical Database. *Int J Lexicography*, vol. 3(4), páginas 235–244, 1990.

MILLER, N. y LEWIS, K. National Survey of Adult Skills in Wales. Informe técnico, Welsh Government Social Research, 2012.

MISHRA, H., MISHRA, A. y SHIV, B. In praise of vagueness: malleability of vague information as a performance booster. *Psychological Science*, vol. 22(6), 2011.

MORO, P., CABERO, M. y RODRÍGUEZ, J. L. Ecuaciones de predicción de lecturabilidad. Informe técnico, e-spacio UNED, 1993.

NEWELL, A. y BOOTH, L. The use of lexical and spelling aids with dyslexics. *Computers and Literacy*, páginas 35–44, 1991.

NIVRE, J. An Efficient Algorithm for Projective Dependency Parsing. En *Proceedings of the 8th International Workshop on Parsing Technologies (IWPT03)*. 2003.

PADRÓ, L., COLLADO, M., REESE, S., LLOBERES, M. y CASTELLÓN, I. 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, 2010.

PETERS, E., HIBBARD, J., SLOVIC, P. y DIECKMANN, N. Numeracy Skill And The Communication, Comprehension, And Use Of Risk-Benefit Information . *Health Affairs*, 2007.

PETERSEN, S. E. y OSTENDORF, M. Text Simplification for Language Learners: A Corpus Analysis. *Speech and Language Technology for Education*, 2007.

PIAGET, J. Essai sur quelques aspects du développement de la notion de partie chez lénfant. *Journal de psychologie normale et pathologique*, vol. 18(6), páginas 449–480, 1921.

PIAGET, J. Une expérience sur le développement de la notion de temps. *Revue suisse de psychologie et de psychologie appliquée*, vol. 1, páginas 179–185, 1942.

PIAGET, J. y INHELDER, B. *Psicología del niño*. Editorial Morata. 1969.

POWER, R. y WILLIAMS, S. Generating Numerical Approximations. *Computational Linguistics*, vol. 38(1), 2012.

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

RAYNER, K. y DUFFY, S. Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition*, vol. 14(3), páginas 191–201, 1986.

REITER, E., SRIPADA, S., HUNTER, J. y YU, J. Choosing words in computer-generated weather forecasts. *Journal Artificial Intelligence*, 2005.

RELLO, L., BAUTISTA, S., BAEZA-YATES, R., GERVÁS, P., HERVÁS, R. y SAGGION, H. One Half or 50 %? An Eye-Tracking Study of Number Representation Readability. En *Proceedings of the 14th IFIP TC13 Conference on Human-Computer Interaction (INTERACT)*. 2013.

SAGGION, H., GÓMEZ-MARTÍNEZ, E., ETAYO, E., ANULA, A. y BOURG, L. Text Simplification in Simplext: Making Text More Accessible. *Procesamiento del Lenguaje Natural*, vol. 47, 2011.

SAQUETE, E., VÁZQUEZ, S., LLORET, E., LLOPIS, F., GÓMEZ, J. y MOSQUERA, A. Tratamiento de textos para mejorar la comprensión lectora en alumnos con deficiencias auditivas. *Procesamiento del Lenguaje Natural*, vol. 50, 2013.

SERENO, S. y RAYNER, K. Measuring word recognition in reading: eye movements and event-related potentials. *Trends in Cognitive Sciences*, vol. 7(11), páginas 489–493, 2003.

SERETAN, V. Acquisition of Syntactic Simplification Rules for French. En *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey, 2012.

SIDDHARTHAN, A. An Architecture for a Text Simplification System. En *Language Engineering Conference*, página 64. IEEE Computer Society, 2002. ISBN 0-7695-1885-0.

SIDDHARTHAN, A. *Syntactic Simplification and Text Cohesion*. Tesis Doctoral, Research on Language and Computation, 2003.

SIDDHARTHAN, A. A survey of research on text simplification. *ITL-International Journal of Applied Linguistics. Special Issue on Readability and Text Simplification. Peeters Publishers, Belgium*, 2014.

SIDDHARTHAN, A. y ANGROSH, M. Hybrid text simplification using synchronous dependency grammars with hand-written and automatically harvested rules. En *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014)*. Gothenburg, Sweden, 2014.

SOLÉ, I. *Estrategias de lectura*. Editorial Graó. 1999. ISBN 9788478278688.

SPAULDING, S. Two Formulas for Estimating the Reading Difficulty of Spanish. *Lawrence Erlbaum Associates, Inc.*, vol. 30(5), páginas 117–124, 1951.

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

UN. Normas Uniformes sobre la igualdad de oportunidades para las personas con discapacidad ONU. Informe técnico, United Nations, 1994.

VELLUTINO, F., FLETCHER, J., SNOWLING, M. y SCANLON, D. Specific reading disability (dyslexia): What have we learned in the past four decades? *Journal of child psychology and psychiatry*, vol. 45(1), páginas 2–40, 2004.

WALKER, A., SIDDHARTHAN, A. y STARKEY, A. Investigation into human preference between common and unambiguos lexical substitutions. En *Proceedings of the 13th European Workshop on Natural Language Generation*. 2011.

WILLIAMS, J., CLEMENS, S., OLEINIKOVA, K. y TARVIN, K. A national needs and impact survey of literacy, numeracy and ICT skills. Informe técnico, 2003.

WILLIAMS, J., CLEMENS, S., OLEINIKOVA, K. y TARVIN, K. Skills for Life Survey 2011. Informe técnico, Department for Business Innovation and Skills. UK, 2012.

WILLIAMS, S. A Parser and Information Extraction System for English Numerical Expressions. Informe técnico, The Open University, Milton Keynes, MK7 6AA, U.K., 2010.

WILLIAMS, S. y POWER, R. Precision and mathematical form in first and subsequent mentions of numerical facts and their relation to document structure. En *Proceedings of the 12th European Workshop on Natural Language Generation*. Athens, 2009.

WILLIAMS, S. y POWER, R. A Fact-aligned Corpus of Numerical Expressions. En *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. Valletta, Malta, 2010.

WILLIAMS, S. y REITER, E. Generating readable texts for readers with low basic skills. En *Proceeding of the 10th European Workshop on Natural Language Generation*, páginas 140–147. Aberdeen, Scotland, 2005.

WILLIAMS, S. y REITER, E. Generating basic skills reports for low-skilled readers. *Journal Natural Language Engineering*, 2008.

WOODSEND, K. y LAPATA, M. Learning to Simplify Sentences with Quasi-Synchronous Grammar and Integer Programming. En *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Edinburh, Scotland, UK, 2011.

WUBBEN, S., VAN DEN BOSCH, A. y KRAHMER, E. Sentence Simplification by Monolingual Machine Translation. En *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*. 2012.

YATSKAR, M., PANG, B., DANESCU-NICULESCU-MIZIL, C. y LEE, L. For the sake of simplicity: Unsupervised extraction of lexical simplifications from Wikipedia. En *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*. 2010.

ZHU, B. D., Z. y GUREVYCH, I. A monolingual tree-based translation model for sentence simplification. En *Proceedings of the 23rd International Conference on Computational Linguistics, COLING'10*. 2010.

Cuando una persona desea realmente algo,
el Universo entero conspira para que pueda realizar su sueño.
Basta con aprender a escuchar los dictados del corazón
y a descifrar un lenguaje que está más allá de las palabras,
el que muestra aquello que los ojos no pueden ver.
El Alquimista
Paulo Coelho