Procesamiento del Lenguaje Natural, Revista nº 49 septiembre de 2012, pp 13-20 recibido 04-05-12 revisado 01-06-12 aceptado 06-06-12

Reducing Text Complexity through Automatic Lexical Simplification: an Empirical Study for Spanish

Reducción de la complejidad de un texto a través de la simplificación léxica: un estudio para el español

Biljana Drndarević y **Horacio Saggion**
Universitat Pompeu Fabra
Departament of Communications and Information Technology
C/Tànger 122, Barcelona
{biljana.drndarevic, horacio.saggion}@upf.edu

Resumen: En este artículo presentamos los resultados de un estudio cuyo objetivo es sentar las bases para el desarrollo de un módulo de simplificación léxica para el español. Basándonos en estudios para otras lenguas analizamos, en primer lugar, la distribución de la frecuencia y la longitud de palabra en textos originales y sus simplificaciones manuales. En segundo lugar nos centramos en los casos de clarificación de información a través de la introducción de definiciones en textos simplificados. Finalmente estudiamos la reducción del contenido informativo del texto y proponemos un sistema para su tratamiento basado en técnicas de resumen. Nuestro estudio empírico sienta las bases para el desarrollo de un componente de tratamiento léxico en un sistema de simplificación de textos en desarrollo.
Palabras clave: simplificación léxica, frecuencia, longitud de palabra, reducción del contenido

Abstract: In this paper we present the results of a study directed towards developing a lexical simplification module of an automatic simplification system for Spanish, intended for readers with cognitive disabilities. We here observe the word length and frequency distribution of two sets of texts that make up our parallel corpus, and we focus on cases of information expansion (through the insertion of definitions)and content reduction (through summarisation). Our ultimate goal is computational implementation of lexical changes in the future.
Keywords: lexical simplification, word frequency, word lengh, information expansion, content reduction

1 Introduction

The digitalisation of information as an essential characteristic of our society has created an illusion of an ideal world where information is freely shared and equally accessible to everyone. Yet, the reality is disappointingly different, as shown by the results of a UN audit conducted with the aim of testing the state of accessibility of 100 leading websites around the world. Only three web pages achieved basic accessibility status. As a result, we have witnessed an increased interest in the issues of e-Accessibility, i.e. the ability for individuals with specific needs to access digital content. For that reason, in recent years NLP has seen a growing number of automatic text simplification systems developed for a wide range of end users. The need and interest for such systems arise from the fact that text is often so complex that it results incomprehensible.

Our project follows this line of research, centering on the development of a tool for automated simplification of newspaper articles in Spanish, meant as an aide for readers with cognitive disabilities. We are currently working on a lexical simplification module, more specifically detecting types of lexical change in a parallel corpus of original and manually simplified news articles, with the aim of preparing their computational implementation. The importance of lexical transformations in text simplification has already been underlined in previous work (Caseli et al., 2009; Specia, 2010). Our corpus analysis has also shown that lexical changes are the most common type of operations carried out by human editors. In broad terms, words and

ISSN 1135-5948

expressions perceived as complicated are substituted with simpler synonyms or rewritten using paraphrase. As shown in previous work (Carroll et al., 1998; Bautista, Gervás, and Madrid, 2009) and the simplification guidelines followed to obtain simplified texts for our corpus[1], "complicated" words tend to be longer and less frequently used ones. Hence, for example, *médico* (*doctor*) is preferred instead of its longer and less frequently used hyponym *psichiatra* (*psychiatrist*). We will, therefore, observe the distribution of word frequency and word length in the original and simplified texts in our corpus with the aim of testing how relevant the combination of these factors might be when conducting synonym substitution. In addition to that, we concentrate on cases of information expansion and content reduction. The former occurs through the insertion of definitions of difficult terms, where, for example, *Amnesty International* is defined as *an organisation that defends human rights worldwide*. On the other hand, content reduction is most often seen with numerical expressions and named entities.

The remainder of this paper is organised as follows: Section 2 addresses the related work in the field of automatic lexical simplification; in Section 3, we describe our methodology, followed by Section 4 where we discuss the results of our study. We conclude and outline our future work in Section 5.

2 Related Work

Previous work in the field of automatic text simplification already established the importance of lexical change. Carroll et al. (1998) presented a project for simplification of news articles in English, in which they used a combination of synonym look-up and word frequency count to carry out lexical substitution. For every content word in the input text, a set of synonyms is extracted from WordNet and Kucera-Francis frequencies are searched in the Oxford Psycholinguistic Database (Quinlan, 1992), upon which the most frequent synonym from the set is chosen for the simplified version of the text. Similar approach has been used in a number of other works. Lal and Ruger (2002) borrowed this method to deal with the lexical component of their automatic text summarizer. Burstein et al. (2007) focused on vocabulary changes when offering ATA V.1.0 as a text adaptation tool for L2 teachers and language learners. Bautista, Gervás, and Madrid (2009) also refer to a thesaurus when extracting candidates for lexical substitution, but their choice is guided by word-length rather than frequency. Caseli et al. (2009) build a parallel corpus for Brazilian Portuguese and extract lexical simplification operations applied by a human annotator, using a list of simple words and a list of discourse markers as resources for synonym substitution.

Acknowledging the fact that many words are polysemic and that, therefore, simple synonym substitution does not always produce a felicitous output, De Belder, Deschacht, and Moens (2010) suggest the use of word sense disambiguation techniques in order to account for contextual information. For every given word they create two sets of "alternative words" – one based on synonyms from WordNet or a similar dictionary, and another one generated by means of the Latent Words Language Model. Once the intersection of these two sets is found, the probability which determines whether it is a suitable replacement is calculated for every word of the intersection. To measure the probability they take into account the difficulty of the word, based on Kucera-Francis frequency, the average number of syllables and unigram probability extracted from a corpus of easy-to-read texts, such as Simple English Wikipedia.

3 Methodology

In order to conduct data analysis we have gathered a corpus of 200 news articles in Spanish. Subsequently, 40 articles have been manually simplified by trained human editors following easy-to-read guidelines proposed by Anula (2009). The most relevant for our current work are preserving the essential information and eliminating any superfluous content; using higher frequency words and avoiding technical terms; and avoiding long words and substituting them with their shorter synonyms with the same frequency index. However, we are interested to see how human editors deal with cases not envisioned by the guidelines, as well as those not described in sufficient detail, such as, for example, what terms are to be explained by means of a definition.

[1]The guidelines are currently in the form of internal project documentation and are to be published at a later date.

Both original and simplified texts have been automatically annotated with FreeLing (Padró et al., 2010). Additionally, sentence alignments have been produced automatically and any errors have been manually corrected through an alignment plug-in in GATE (Cunningham et al., 2002), a graphical editing tool for text processing. We have observed cases where one original (O) sentence corresponds to one or more simplified (S) sentences, cases where the relation is reversed as well as cases where there is no correlation between O and S sentences, due to information elimination or expansion.

We analyse thus aligned pairs of O and S texts in order to detect simplification operations applied at the lexical level, upon which we ponder the possibility of applying these operations computationally. In addition to that, we conduct text processing at the word level in order to gather data relative to word frequency and word length. Previous work having mainly concentrated on word frequency when applying synonym substitution, our intention is to test on our parallel corpus how this factor combines with that of word length, a traditional readability metric. Frequencies are extracted from a dictionary based on the Referential Corpus of Contemporary Spanish[2]. Every word in the dictionary is assigned a relative frequency index (FI) from 1 to 6, where 1 represents the lowest frequency and 6 the highest. The words that do not appear in the dictionary are assigned FI 0. We placed these words in three different categories: named entities, numerical expressions and what we call rare words. Among rare words we encounter multi-word expressions, such as complex function words, like *a través de* (*by means of*). This is due to the fact that multi-word expressions are recognized as such by FreeLing, whereas the current version of the frequency dictionary does not contain such words. However, the ratio of these words with respect to the total is fairly small so as to significantly influence overall results (1.08% in O and 0.59% in S).

4 Data Analysis

The corpus analysis has provided us with an insight into what lexical elements are treated and in what manner. We here concentrate on the insertion of definitions of difficult terms and concepts, and the treatment of named entities (NE) and numerical expressions (NumExp). In general, these can be seen as cases of information expansion on the one hand (insertion of definitions), and information elimination on the other, since a significant number of NE and NumExp are eliminated. In addition to that, we processed all original and simplified texts, placed into two separate sets (O and S), with the aim of obtaining a quantitative description of these sets, as reflected in the following:

- average sentence length;
- average number of sentences per text;
- average word length (in characters);
- the distribution of n-character words;
- the distribution of n-frequency words.

Sections that follow summarise the results of the quantitative analysis of O and S text sets, the treatment of definitions, named entities and numerical expressions.

4.1 Word length and frequency

Table 1 summarizes the data relative to text, sentence and word length, where s/t stands for "sentence per text" while w/s represents "word per sentence".

	Original	**Simple**
Total words	6595	3912
Total sentences	246	324
Average s/t	6.64	8.75
Average w/s	26.8	12.07
Average word length	5.44	5.07

Table 1: Average text, sentence and word length in original and simplified texts

As can be appreciated, S texts tend to be quite shorter on the whole, containing around 40% fewer words than O texts. However, they contain 24% more sentences than O texts, and their sentences are more than 50% shorter. The tendency is clear – long O sentences are generally split into shorter ones, and a considerable amount of O content is eliminated. We will explore the latter in more detail in Section 4.3.

One curious observation is that relative to average word length – contrary to our expectations, S words are only slightly shorter than O ones. We therefore focused on all

[2]http://corpus.rae.es/creanet.html

words with 1, 2, ... 20 characters, while longer words have been placed in categories of words with 21-30 characters, words with 31-40 characters and words with more than 40 characters.[3] The data analysis revealed that the most prolific words in both O and S texts are two-character words, accounting for as much as 27% of the texts. The vast majority of these are function words (97.61% in O and 88.97% in S). We have also observed that three to seven-character words are more abundant in S texts, whereas longer words are slightly more common in O texts. However, it is interesting to note that S texts contained on average slightly more eighteen-character words and words containing between 21 and 30 characters. These are all cases of named entities, often repeated through the insertion of definitions (discussed more in depth in Section 4.4).

On the whole, we can conclude that in S texts there is a tendency towards using shorter words of up to ten characters, with one to five-character words taking up 59.81% of the set and one to ten-character words accounting for 94.04% of the content. Longer words are almost exclusively reserved for named entities, often repeated when a definition of the terms in question is inserted.

Apart from word length, we explored how frequency acts as a factor in distinguishing between original and simplified, or difficult and easy words. We have detected words with frequency index 3, 4, 5 and 6, as well as words absent from the dictionary and were therefore assigned FI 0. The latter include numerical expressions (NumExp), named entities (NE), and what we here call *rare words*, i.e. all words not found in the dictionary that are neither NE nor NumExp. A small number of these are multi-word expressions, but the majority are indeed words not used very often, such as *intransigencia* (*intransigence*) or foreign words, like *e-book*. Table 2 contains data relative to average number of n-frequency words in O and S texts, where zero frequency words have been additionally separated into the categories of NumExp, NE and rare words and are printed in bold.

We can observe that the frequency distribution is fairly equal in both sets of texts, with the greatest divergence in the category of rare words, which are more than 50% less abundant in S texts than O texts. NE are slightly more common in S texts, due to the fact that these are often repeated - we have observed a preference for using NE instead of referring expressions like pronouns or definite noun phrases in S texts (see Section 4.2). We should also acknowledge that low frequency words (FI 3) are used around half as much in S as in O texts, while the former is somewhat more saturated in words with the highest frequency rate, in line with our predictions.

Frequency index	Original	Simple
Rare words	9.49%	4.19%
NE	7.08%	8.77%
NumExp	2.81%	2.02%
Freq. 0 total	19.38%	14.98%
Freq. 3	1.23%	0.66%
Freq. 4	1.21%	0.89%
Freq. 5	6.02%	5.06%
Freq. 6	72.16%	78.40%

Table 2: The distribution of n-frequency words in original and simplified texts

If, additionally, we analyse the word length of rare words, we notice that the majority of these (72.44% in O and 77.44% in S) are words made up of seven to nine characters, followed in percentage by longer words of up to twenty characters in O texts (39.42%) and fourteen characters in S texts (29.88%). We could, therefore, draw a general conclusion that longer words tend to be used more sparingly in S texts and that the combination of factors such as frequency and word length might be the one to be taken into account when carrying out lexical substitution based on synonymy.

4.2 Named Entities and Numerical Expressions

Examining the parallel corpus, we have observed that NumExp and NE are given special attention when simplifying texts for people with cognitive disabilities. We have documented numerous cases of such expressions, as well as changes applied to them. One common operation is the substitution of a definite noun phrase with a NE it refers to. For example, *the Andalusian town* is substituted with *Granada*. As for NumExp, a good example of common simplification operations are

[3] Words treated here are the result of processing with FreeLing, where multi-word expressions, among them named entities or numerical expressions, are treated as single words – hence words of more than 20 characters in our corpus.

rounding of big numbers, eliminating NumExp from parenthesis and the use of numerical modifiers, such as *almost* or *more than*, all three illustrated in the following pair of original (1) and simplified (2) sentences:

1. *The Secretary General of the UN, Ban Kimoon, asked for major funding for humanitarian actions in 2011, with a petition of* ***almost 7,400 million dollars*** *(****around 5,400 million euros****).*
2. *The Secretary General of the UN asked for* ***more than 7,000 million dollars*** *for humanitarian actions.*

However, our data shows that by far the most common operation applied to NumExp and NE is elimination. Almost 60% of the original NumExp have been eliminated as a result of simplification. In the case of NE, the average number of NE in simplified texts is slightly higher than in original ones (8.77% in S and 7.08% in O). This, however, is due to the fact that NE seen as essential for the core message of the text are often repeated, both through the introduction of definitions of such terms and the use of NE instead of a definite noun phrase, as already seen. When the number of *different* NE are counted in each set, we perceive a strong tendency towards elimination – S texts contain half as many NE types as O texts. The following sentences illustrate a case of NE elimination (the eliminated expressions are printed in bold).

1. *Today the Mayor of* ***Madrid****,* ***Alberto Ruiz-Gallardón****, inaugurated the new library, situated in the* ***Cultural Centre Eduardo Úrculo*** *and dedicated to the philosopher* ***María Zambrano****; the library caters for six neighbourhoods in the district of Tetuán.*
2. *The new library is in the Tetuán district.*

4.3 Sentence Elimination

Content reduction in text simplification is not only observed in the elimination of certain phrases such as NumExp or NE, as already suggested, but also in the deletion of full sentences. As our corpus study indicates, 20% of all sentences in O texts are deleted to create S texts. Even though one could argue that this percentage is too small to justify implementing a deletion operation, it is a striking fact that 72% of O texts in our corpus contain at least one case of sentence elimination. Therefore, we argue that a sentence deletion procedure should be a key element in making texts simpler, since it is indeed a very productive operation. The module to simplify content through sentence deletion is implemented as a sentence classification mechanism: it decides which sentences from O texts to delete, the data for training the classifier being the set of all O sentences annotated with a feature indicating whether the sentence should be deleted or kept. Every sentence in the corpus is represented as a set of features, some of them borrowed from text summarisation and others specific to our problem. For example, we consider that the position of the sentence in the text may be a factor when deciding whether to delete it or not. In fact, in the informative discourse we are treating, less "topical" sentences would likely appear towards the end of the document, being therefore good candidates for deletion. Other features we are considering are the number of NE and NumExp in the sentence (justified by our corpus study), the number of content words in the sentence, and the number of punctuation tokens. Various cohesion features are computed as the number of shared content words units between neighbouring sentences: this is done to implement topic shifts. Word frequency distribution is also used as a feature. Average word frequency is calculated for every sentence and this information is used as one of the features for the classifier. The classification system is based on a Support Vector Machines implementation (Li et al., 2002) that can be used for training, testing, and cross-validation experiments. We have considered two simple baseline (non-trainable) procedures which delete the last or last two sentences of each document (See Table 3). One of the baselines already provides a very reasonable performance with an F-score (F1)of 0.73. However, our more informed classifier, trained with our designed features, reaches an improved F-score of 0.79 in cross-validation experiments, improving on both precision and recall of the two baselines. The classifier performance needs to be improved, especially for recognising delete cases.

4.4 Insertion of Definitions

In 57.5% of all texts we found cases of S sentences with no correlation to O sen-

	Delete			Keep			Overall
Condition	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**F1**
Delete last	0.27	0.20	0.23	0.81	0.86	0.84	0.73
Delete 2 last	0.31	0.46	0.37	0.84	0.74	0.79	0.68
Classifier	0.42	0.26	0.30	0.86	0.89	0.87	0.79

Table 3: Results of Cross-validation Sentence Deletion Experiments: Baselines and Classifier

tences. These are all cases of definitions of difficult terms and concepts, inserted in the text as additional information. The majority of these are definitions of named entities, such as personal names (*El Greco*), organizations (*the United Nations*), geographical terms (*Guantanamo*) and alike. A certain number of lexical units are also explained by means of a definition, 80% of which are zero frequency words. Hence, for example, *molecules* are defined as *very small parts of the universe*.

As already shown in previous sections, both named entities and rare words are perceived as complicated. In the majority of cases, such terms are either eliminated or replaced by their synonyms with higher FI. However, when these terms are central to the core theme of the text, they cannot be eliminated. Synonym substitution is not always possible either, since NE do not have synonyms and nor do extremely rare and technical terms (like *molecules* from the above example). This is where definitions are inserted, as a means of simplifying complicated but essential elements of information.

In an attempt to investigate the issue of definition insertion as a possible component of our lexical simplification module, we catalogued all such expressions from the corpus[4] and juxtaposed them to definitions extracted from web sources. We then conducted quantitative analysis of both sets of definitions, termed *long* and *short* to avoid confusion with original and simplified texts. The following pair of sentences are an example of a long definition (1), taken from the web, and a short definition (2) inserted by human editors:

1. *Alhambra is a monumental complex created over the period of more than six hundred years by such diverse cultures as the Muslim, the Renaissance and the Romance culture.*
2. *Alhambra is an Arabic monument in Granada.*

[4]Definitions from the corpus were created from scratch by trained human editors.

We can see that not only does the long definition contain a lot more words, but some of the words are among less frequently used ones, such as *monumental complex* or *the Renaissance*. It is, therefore, clear that simple insertion of definitions found on the web or in encyclopaedias does not necessarily contribute to creating a "simple" text - further simplification of inserted sentences is necessary. In order to test that hypothesis, we analysed average word length and frequency distribution in both sets of definitions. Table 4 provides the obtained figures.

	Long	**Short**
Word length	5.80	2.17
Sentence length	27.74	11.37

Table 4: Word length and sentence length in long and short definitions

As can be appreciated, there is a significant discrepancy in both word and sentence length between short definitions and the ones found on the web. Sentences in short definitions are more than half as long as the ones found in long definitions, and a strong preference for the use of short words is observed in short definitions.

As for frequency distribution, presented in Table 5, we notice a similar pattern as when comparing O and S texts: the majority of the words are high frequency words, whereas the rate of low frequency words is rather negligible. Where the two sets do differ more significantly is the distribution of zero frequency words. The percentage of actual rare words (i.e. not NumExp and NE) is significantly less common in short definitions than in long ones - the latter contain four times as many rare words. NumExp are fairly rare in both sets, with the short set containing only one such example. What we mean by *defined terms* are those terms for which the definition is being inserted, like *Amnesty International* or *molecules*. Since the vast majority of the defined terms are NE, we placed them to-

FI	Long	Short
NumExp	1.61%	0.46%
Defined terms and other NE	6.66%	12.04%
Rare words	9.87%	2.31%
Freq. 0 total	18.14 %	14.81%
freq. 3	0.89%	1.39%
freq. 4	1.61%	0.46%
freq. 5	6.60%	3.70%
freq. 6	72.77%	79.63%

Table 5: The distribution of n-frequency words in long and short definitions

gether in the category of "defined terms and other NE". A somewhat striking initial observation is that such terms are twice as common in the short definition set as in the long one. In order to further analyse the distribution of this category of words in both sets, we divided the set into "defined terms proper", which include both NE and lexical units, and "other NE", which include NE other than the ones being defined. Subsequently, we calculated the number of different NE among the category "other NE", in order to see how repetition influences the number of these words in the definitions. Table 6 summarises the percentages of *defined terms*, *other NE* and *different NE* against the total number of *NE and defined terms*.

Word category	Long	Short
Defined terms	46.67%	80.77%
Other NE	53.33%	19.23%
Different NE	38.33%	19.23%

Table 6: Percentage of defined terms and other named entities in long and short definitions

As can appreciated, there is a stark difference between the two sets – there is more repetition of defined terms in short definitions, reducing the introduction of new named entities to the minimum (only five NE in total, with four different NE). In long definitions, however, the two categories of words are balanced out. In addition to that, the total of 65 long definitions contains 46 different named entities, which is in terms of percentages almost double the number of different NE in the short definition set. The following pair of sentences, introducing the definition of "Congo", are an illustration:

1. ***Democratic Republic of Congo**, previously known as **Zaire** and in the colonial period as **Belgian Congo**, is a country in **Africa**, with the capital in **Kinshasa**.*
2. ***Congo** is a country in **Africa**.*

As can be observed, the long definition (1) uses five different NE, introducing four new ones and changing the form of the NE being defined. On the other hand, the short definition (2) only introduces one extra NE and leaves the defined term unchanged.

Based on the figures analysed above we could draw the following conclusions:

- Definitions should employ short words with higher frequency index.
- Rare words (FI 3 or below) should be avoided.
- The introduction of NE other than those being defined should be avoided.
- A NE or a term being defined may be repeated in order to underline it and allow the reader to memorise it.

As part of our future work we intend to further explore the issue of deciding which terms to define, which to eliminate and where to apply synonym substitution.

5 Conclusions and Future Work

In this paper we presented the results of a quantitative analysis of a parallel corpus of original and manually simplified texts in Spanish, conducted with the aim of observing how word length and frequency act as factors to determine word difficulty and influence the choice of synonyms selection when applying lexical substitution. We have found that almost 95% of the words in simplified texts consist of up to ten characters, whereas original texts contain a larger number of longer words. As for frequency, simplified texts tend to contain a slightly greater number of high frequency words, whereas what we call *rare words* are almost 50% more common in original texts. Additional analysis shows that these words tend to be up to 20 characters long in original and 14 characters long in simplified texts. We could, therefore, conclude that the combination of the factors of word length and frequency could be indicative of word difficulty and could be chosen to guide the process of synonym substitution.

We additionally analysed cases of content expansion through the introduction of definitions of complicated terms, mostly named entities. Our results show that definitions are to be composed of short words with higher frequency index and that introduction of new named entities is to be avoided. Therefore, further simplification of definitions found on the web is to be applied for a truly simplified output. On the other hand, we observed the cases of information elimination, with special attention to numerical expressions and named entities, and found that entire sentences can be safely removed in order to produce a more readable text.

As part of our future work, we intend to further investigate the insertion of definitions, focusing on problems such as what terms to define, what to eliminate and where to apply synonyms substitution. Similarly, the performance of the classifier to recognise delete cases is also to be improved. Our final aim is computational implementation of a lexical simplification module as part of a system for automatic text simplification in Spanish, aimed at readers with cognitive disabilities.

Acknowledgements

We present this work as part of a project entitled Simplext: An automatic system for text simplification, with the file number TSI-020302-2010-84 (`http://www.simplext.es`). We are also grateful to the fellowship RYC-2009-04291 from Programa Ramón y Cajal 2009, Ministerio de Economía y Competitividad, Secretaría de Estado de Investigación, Desarrollo e Innovación, Spain.

References

Anula, A. 2009. Tipos de textos, complejidad lingüística y facilicitación lectora. In *Actas del Sexto Congreso de Hispanistas de Asia*, pages 45–61.

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

Burstein, J., J. Shore, J. Sabatini, Yong-Won Lee, and M. Ventura. 2007. The automated text adaptation tool. In *HLT-NAACL (Demonstrations)*, pages 3–4.

Carroll, J., G. Minnen, Y. Canning, S. Devlin, and J. Tait. 1998. Practical simplification of english newspaper text to assist aphasic readers. In *Proc. of AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, pages 7–10.

Caseli, H. M., T. F. Pereira, L. Specia, Thiago A. S. Pardo, C. Gasperin, and S. M. Aluísio. 2009. Building a brazilian portuguese parallel corpus of original and simplified texts. In *10th Conference on Intelligent Text PRocessing and Computational Linguistics (CICLing 2009)*.

Cunningham, H., D. Maynard, K. Bontcheva, and V. Tablan. 2002. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*.

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

Lal, P. and S. Ruger. 2002. Extract-based summarization with simplification. In *Proceedings of the ACL 2002 Automatic Summarization / DUC 2002 Workshop*.

Li, Y., H. Zaragoza, R. Herbrich, J. Shawe-Taylor, and J. Kandola. 2002. The Perceptron Algorithm with Uneven Margins. In *Proceedings of the 9th International Conference on Machine Learning (ICML-2002)*, pages 379–386.

Padró, Ll., M. Collado, S. Reese, M. Lloberes, and I. Castellón. 2010. Freeling 2.1: Five years of open-source language processing tools. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta.

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

Specia, Lucia. 2010. Translating from complex to simplified sentences. In *Proceedings of the 9th international conference on Computational Processing of the Portuguese Language*, pages 30–39, Berlin, Heidelberg.