Towards Automatic Lexical Simplification in Spanish: An Empirical Study

Biljana Drndarević and **Horacio Saggion**
Universitat Pompeu Fabra
Department of Information and Communication Technologies
C/ Tanger, 122-140
08018 Barcelona, Spain
{biljana.drndarevic,horacio.saggion}@upf.edu

Abstract

In this paper we present the results of the analysis of a parallel corpus of original and simplified texts in Spanish, gathered for the purpose of developing an automatic simplification system for this language. The system is intended for individuals with cognitive disabilities who experience difficulties reading and interpreting informative texts. We here concentrate on lexical simplification operations applied by human editors on the basis of which we derive a set of rules to be implemented automatically. We have so far addressed the issue of lexical units substitution, with special attention to reporting verbs and adjectives of nationality; insertion of definitions; simplification of numerical expressions; and simplification of named entities.

1 Introduction

In the highly digitalized 21st century sharing information via Internet has become not only commonplace but also essential. Yet, there are still a large number of people who are denied this fundamental human right – access to information. In 2006 the UN conducted an audit with the aim of testing the state of accessibility of the leading websites around the world. The results were rather disappointing, with only three out of 100 tested web pages achieving basic accessibility status. It is therefore clear that one of the priorities for the future is working on enabling inclusion of all the groups that are currently marginalised and denied equal access to information as the rest of the population.

Written information available online is far too often presented in a way that is perceived as incomprehensible to individuals with cognitive disabilities. It is therefore necessary to simplify the complex textual content in order to make it more accessible. However, manual simplification is too time-consuming and little cost-effective so as to yield sufficient amount of simplified reading material in a satisfactory time frame. Hence, the need and interest arise to develop automatic or semi-automatic simplification tools that would (partially) substitute humans in carrying out this laborious task.

Our project is one such aspiration. Our goal is to offer an automated text simplification tool for Spanish, targeted at readers with cognitive disabilities. We delimit our research to simplification of informative texts and news articles. So far we have focused primarily on syntactic simplification, with an already implemented module currently in the test stage (Bott and Saggion, 2012b). The present work, however, deals with lexical simplification and is centred around a corpus analysis, a preparatory stage for the development of a separate lexical module in the future.

Earlier work already establishes the importance of lexical changes for text simplification (Carroll et al., 1998; Caseli et al., 2009; De Belder et al., 2010). Upon examining a parallel corpus consisting of original and manually simplified newspaper articles in Spanish, we have found that by far the most common type of changes applied by human editors are precisely lexical changes, accounting for 17.48% of all annotated operations (Bott and Saggion, 2012a). Words perceived as more complicated are replaced

NAACL-HLT 2012 Workshop on Predicting and Improving Text Readability for target reader populations (PITR 2012)., pages 8–16,
Montréal, Canada, June 7, 2012. ©2012 Association for Computational Linguistics

by their simpler synonyms. A recurring example is that of reporting verbs. Corpus analysis shows a clear tendency towards replacing all reporting verbs such as *advertir* (*warn*), *afirmar* (*declare*), *explicar* (*explain*), etc. with the ubiquitous *decir* (*say*). Sentences 1 (original) and 2 (simplified) illustrate the said phenomenon (translated into English):

1. *It is important that we continue working on the means that promote the access of the disabled to cultural content, she* ***explained****.*

2. *The Minister of Culture* ***said*** *that she is working towards granting the disabled access to cultural content.*

We therefore document all cases of lexical change observed in the corpus and try to extract rules for their automatic implementation. The remainder of this paper is organized as follows: Section 2 addresses the related work in the field; in Section 3 we describe the experimental setting and the process of obtaining the parallel corpus used in the study, while Section 4 provides a more detailed insight into the kind of lexical simplifications observed. We conclude in Section 5 and outline our future work.

2 Related Work

Text simplification has so far been approached with two different aims. One is to offer simplified versions of original text to human readers, such as foreign language learners (Petersen and Ostendorf, 2007; Medero and Ostendorf, 2011); aphasic people (Devlin and Unthank, 2006); low literacy individuals (Specia, 2010) and others. On the other hand, simplified text is seen as input for further natural language processing to enhance its proficiency, e.g. in machine translation or information retrieval tools (Klebanov et al., 2004). The earliest simplification systems employed a rule-based approach and focused on syntactic structure of the text (Chandrasekar et al., 1996). The PSET project (Carroll et al., 1998) dealt with simplification of news articles in English for aphasic readers. Together with syntactic analysis and transformations similar to those of Chandrasekar et al. (1996), they employed lexical simplification based on looking up synonyms in WordNet and extracting Kucera-Francis frequency from the Oxford Psycholinguistic Database (Quinlan, 1992). Therefore, the most frequent of a set of synonyms for every content word of the input text was chosen to appear in its simplified version.

The above approach to lexical simplification has been repeated in a number of works (Lal and Ruger, 2002; Burstein et al., 2007). Bautista et al. (2009) also rely on a dictionary of synonyms, but their criterion for choosing the most appropriate one is word-length rather than frequency. Caseli et al. (2009) analyse lexical operations on a parallel corpus of original and manually simplified texts in Portuguese, using lists of simple words and discourse markers as resources. Bautista et al. (2011) focused on numerical expressions as one particular problem of lexical simplification and suggested the use of hedges as a means of dealing with complex numerical content.

Given the fact that many words tend to be polysemic, attempts have been made to address this issue so as to provide more accurate, context-aware lexical substitution. De Belder et al. (2010) were the first to employ word sense disambiguation techniques in order to capture contextual information, while Biran et al. (2011) apply an unsupervised method for learning pairs of complex and simple synonyms based on an unaligned corpus of texts from the original Wikipedia and Simple English Wikipedia.

3 Experimental Setting

We have gathered a corpus consisting of 200 informative texts in Spanish, obtained from the news agency Servimedia. The articles have been classified into four categories: national news, international news, society and culture. We then obtained simplified versions of the said texts, courtesy of the DILES (Discurso y Lengua Española) group of the Autonomous University of Madrid. Simplifications have been applied manually, by trained human editors, following easy-to-read guidelines suggested by Anula (2009), (2008). We are interested to see how these guidelines are applied in practice, as well as how human editors naturally deal with cases not treated by the guidelines in sufficient detail.

The corpus has been automatically annotated using part-of-speech tagging, named entity recognition and parsing (Padró et al., 2010). Furthermore, a text aligning algorithm based on Hidden Markov Models (Bott and Saggion, 2011) has been applied to obtain sentence-level alignments. The automatic alignments have then been manually corrected through a graphical editing tool within the GATE framework (Cunningham et al., 2002). A total of 570 sentences have been aligned (246 in original and 324 in simple texts), with the following correlations between them: *one to one*, *one to many* or *many to one*, as well as cases where there is no correlation (cases of content reduction through summarisation or information expansion through the introduction of definitions). The alignments facilitate the observation of the corpus, particularly cases where entire sentences have been eliminated or inserted.

A parallel corpus thus aligned enables us to engage in data analysis as well as possibly carry out machine learning experiments to treat specific problems we have so far detected. We have documented all simplification operations used by human editors and placed them in eight major categories applied at various linguistic levels (individual words, phrases or sentences). The operations are *change, delete, insert, split, proximization, re-order, select* and *join*, listed in the decreasing order of their relative frequency in the corpus. Among these are the changes that are either rather idiosyncratic or involve complex inferential processes proper to humans but not machines. Sentence 1 (original) and paragraph 2 (simplified) are an example (translated into English):

1. *Around 390,000 people have returned to their homes after being forced to evacuate due to floods caused by monsoon rains last summer in Pakistan.*

2. *Last summer it rained a lot in Pakistan.* ***The rain flooded the fields and the houses. That is to say, the water covered the houses and the fields.*** *For this reason a lot of people left their homes in Pakistan. Now these people return to their homes.*

Sentences in bold are examples of information expansion which is difficult to implement automatically. The concept of flood is obviously perceived as complicated. However, instead of offering a definition taken out of a dictionary and applicable to any context (as in the example further below), the writer explains what happened in this particular instance, relying on their common knowledge and inferential thinking. It is obvious that such conclusions cannot be drawn by computers. What *can* be done is insert a definition of a difficult term, as in the following example:

1. *The Red Cross asks for almost one million euros for the 500,000 Vietnamese affected by the floods.*

2. *The Red Cross asks for one million euros for Vietnam.* ***The Red Cross is an organization that helps people and countries around the world.***

After documenting all the operations and analysing their nature and frequency, we have finally decided to focus on the automatic treatment of the following: lexical simplification, deletions, split operations, inversion of direct speech and the insertion of definitions. In the next section, we concentrate on operations applied at the lexical level, with the aim of drawing conclusions about the nature of lexical simplification carried out by trained editors and the possibility of their automatic implementation in the future.

4 Data Analysis

We have so far obtained forty simplifications and our goal is to shortly acquire simplified versions of all 200 texts. A variety of lexical operations have been observed in the corpus, which go far beyond simple substitution of one lexical unit with its simpler equivalent. In order to describe the nature of these changes, we have categorized them as follows:

- substitutions of one lexical unit with its simpler synonym;
- insertion of definitions of difficult terms and concepts;
- simplification of numerical expressions;
- simplification of named entities;
- elimination of nominalisation;
- rewording of idioms and collocations; and
- rewording of metaphorically used expressions.

4.1 Lexical substitution

We have documented 84 cases where one lexical unit has been substituted with its simpler synonym. These words make up our lexical substitution table (LST), gathered for the purpose of data analysis. The table contains the lemma of the *original* (O) word, its *simple* (S) equivalent and additional information about either the original word, the simple word or the nature of the simplification, such as *polysemy*, *hyponym* ⇒ *hypernym*, *metaphor*, etc. Table 1 is an excerpt.

Original	Simple	Commentary
impartir	pronunciar	polysemy
informar	decir	reporting verb
inmigrante	extranjero	hyponym ⇒ hypernym
letras	literatura	polysemy

Table 1: An excerpt from the Lexical Substitution Table

To analyse the relationship between the sets of O-S words, we have concentrated on their frequency of use and length (both in characters and syllables).

4.1.1 Word frequency

For every word in the LST, we consulted its frequency in a dictionary developed for the purposes of our project by the DILES group and based on the Referential Corpus of Contemporary Spanish (Corpus de Referencia del Español Actual, CREA)[1]. We have found that for 54.76% of the words, the frequency of the simple word is higher than the frequency of its original equivalent; in 30.95% of the cases, the frequency is the same; only 3.57% of the simple words have lower frequency than the corresponding original ones; and in 10.71% of the cases it was impossible to analyse the frequency since the original word was a multi-word expression not included in the dictionary, as is the case with complex conjunctions like *sin embargo* (*however*) or *pese a* (*despite*).

As can be appreciated, in a high number of cases O and S words have the same frequency of use according to CREA. In an intent to rationalise this phenomenon, we have counted the number of times each of these words appears in the totality of original and simple texts. In more than half of the O-S pairs the simple word is more common than its original equivalent, not only in the simplified texts, where it is expected to abound, but also in the original ones. This difference in the frequency of use in actual texts and the CREA database could be explained by the specificity of the genre of the texts in our corpus, where certain words are expected to be recurrent, and the genre-neutral language of CREA on the other hand. Out of the remaining 44.5% of the cases, where O words are more abundant than S words, five out of fourteen may have been used for stylistic purposes. One good example is the use of varied reporting verbs, such as *afirmar* (*confirm*) or *anunciar* (*announce*), instead of uniformly using *decir* (*say*). Six in fourteen of the same group are polysemic words possibly used in contexts other than the one where the simplification was recorded. Such is the example of the word *artículo*, substituted with *cosa* where it meant *thing*. However, it also occurs with its second meaning (*article: a piece of writing*) where it cannot be substituted with *cosa*.

What can be concluded so far is that frequency is a relatively good indicator of the word difficulty, albeit not the only one, as seen by a large number of cases when the pairs of O-S words have the same frequency. For that reason we analyse word length in Section 4.1.2. Polysemy and style are also seen as important factors at the time of deciding on the choice of the synonym to replace a difficult word. Whereas style is a factor we currently do not intend to treat computationally, we cannot but recognize the impact that polysemy has on the quality and accuracy of the output text. Consider the example of another pair of words in our lexical substitution table: *impresión* ⇒ *influencia*, in the following pair of original (1) and simplified (2) sentences:

1. Su propia sede ya da testimonio de la “impresión profunda” que la ciudad andaluza dejó en el pintor.
   Its very setting testifies to the profound influence of the Andalusian town on the painter.

2. En esa casa también se ve la influencia de Granada.
   The influence of Granada is also visible in that house.

In the given context, the two words are perfect syn-

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

onyms. However, in expressions such as *tengo la impresión que* (*I am under the impression that*), the word *impresión* cannot be substituted with *influencia*. We have found that around 35% of all the original words in the LST are polysemic. We therefore believe it is necessary to include a word sense disambiguation approach as part of the lexical simplification component of our system in the future.

4.1.2 Word Length

Table 2 summarizes the findings relative to the word length of the original and simple words in the LST, where *syll.* stands for *syllable* and *char.* for *character*.

Type of relationship	Percentage
S has fewer syll. than O	57.85%
S has more syll. than O	17.85%
S has the same number of syll. as O	25%
S has fewer char. than O	66.66%
S has more char. than O	23.8%
S has the same number of char. as O	9.52%

Table 2: Word length of original and simple words

The average word length in the totality of original texts is 4.81 characters, while the simplified texts contain words of average length of 4.76 characters. We have also found that the original and simplified texts have roughly the same number of short words (up to 5 characters) and medium length words (6-10 characters), while the original texts are more saturated in long words (more than 11 characters) than the simplified ones (5.91% in original and 3.64% in simplified texts). Going back to the words from the LST which had the same frequency according to CREA, we found that around 80% of these were pairs where the simple word had fewer syllables than the original one. This leads us to the conclusion that there is a strong preference for shorter words and that word length is to be combined with frequency when deciding among a set of possible synonyms to replace a difficult word.

4.2 Transformation rules

Upon close observation of our data, we have derived a set of preliminary simplification rules that apply to lexical units substitution. These rules concern reporting verbs and adjectives of nationality, and will be addressed in that order.

In the twenty pairs of aligned texts nine different **reporting verbs** are used. All nine of them have been substituted with *decir* (*say*) at least once, amounting to eleven instances of such substitutions. Three verbs from the same set appear in simplified texts without change. On the whole, we perceive a strong tendency towards using a simple verb like *say* when reporting direct speech. Our intention is to build a lexicon of reporting verbs in Spanish and complement it with grammatical rules so as to enable accurate lexical substitution of these items of vocabulary. Simple substitution of one lexical unit with another is not always possible due to syntactic constraints, as illustrated in the following example:

1. El juez advirtió al duque que podría provocar la citación de la Infanta.
   The judge warned the Duke that he might cause the Princess to be subpoenaed.

2. Murió científico que advirtió sobre deterioro de la capa de ozono.
   The scientist who warned about the deterioration of the ozone layer died.

In the first case the verb *advertir* is used as part of the structure [advertir a X que], in English [warn somebody that]. The verb *decir* easily fits this structure without disturbing the grammaticality of the sentence. In the second instance, however, the reporting verb is used with the preposition and an indirect object, a structure where the insertion of *decir* would be fatal for the grammaticality of the output. We believe that the implementation of this rule would be a worthwhile effort, given that informative texts often abound in direct speech that could be relatively easily simplified so as to enhance readability.

As for **adjectives of nationality**, we have noticed a strong preference for the use of periphrastic structure instead of denominal adjective denoting nationality. Thus, a simple adjective is replaced with the construction [de < COUNTRY >], e.g. *el gobierno pakistaní* (*the Pakistani government*) is replaced by *el gobierno de Pakistán* (*the government of Pakistan*). The same rule is applied to instances of nominalised nationality adjectives. In these cases the structure [ArtDef + Adj][2] be-

[2] ArtDef: definite article, Adj: adjective

comes [ArtDef + persona + de + < COUNTRY >], e.g: *los pakistaníes* ⇒ *las personas de Pakistán* (*the Pakistani* ⇒ *the people from Pakistan*). In only five instances the adjective was preferred. Twice it was *español* (*Spanish*), which were the only two instances of the expression of this nationality. This leads us to the conclusion that *español* is sufficiently widespread and therefore simple enough and would not need to be substituted with its periphrastic equivalent. *Norteamericano* (*North American*) was used twice, therefore being slightly more acceptable than *estadounidense* (*of/from the United States*), which is always replaced by *de Estados Unidos*. The remaining is the one instance of *egipcio* (*Egyptian*), otherwise replaced by *de Egipto*.

Based on the observations, our hypothesis is that more common nationality adjectives, such as *Spanish*, and possibly also *English* or *French* need not be modified. *Norteamericano* or *estadounidense* however common are possibly perceived as complicated due to their length. In order to derive a definite rule, we would need to carry out a more detailed analysis on a richer corpus to determine how frequency of use and length of these adjectives correlate.

4.3 Insertion of definitions

Definitions of difficult terms are found in 57.5% of all texts we have analysed. Around 70% of these are definitions of named entities, such as *El Greco*, *Amnesty International*, *Guantanamo* and others. In addition to these, difficult lexical units, and even expressions, are explained by means of a definition. Thus, *a (prison) cell* is defined as *a room in a prison*, and *the prisoner of conscience* as *a person put in prison for his ideas*. In order to deal with named entity definitions, we intend to investigate the methods for the look-up of such definitions in the future. To solve the problem of defining difficult individual lexical units, one solution is to target those words with the lowest frequency rate and in the absence of an adequate simpler synonym insert a definition from a monolingual dictionary, given the availability of such resources (the definition itself might need to be simplified).

4.4 Numerical expressions

Our analysis shows that the treatment of numerical expressions should have a significant place in our simplification system, given their abundance in the kind of texts our system is mainly intended for, and a wide variety of simplification solutions observed by examining the parallel corpus. Even though by far the most common operation is elimination (in the process of summarization), there are a number of other recurrent operations. The most common of these are explained below for the purpose of illustration, given that the totality of the rules is beyond the scope of this paper. We separately address numerical expressions forming part of a date and other instances of using numbers and numerals.

The following are the rules concerning numerical expressions in dates:

1. en < YEAR > ⇒ en el año < YEAR >
   en 2010 ⇒ en el año 2010

2. Years in parenthesis are eliminated (this operation has been applied in 100% of the cases during manual simplification):
   El Greco (1541–1614) ⇒ El Greco

3. In expressions containing the name and/or the day of the month, irrespective of whether it is followed by a year, the information relative to the month (i.e. name or name and day) is eliminated (applied in around 85% of the cases):
   en septiembre 2010 ⇒ en el año 2010
   el 3 de mayo ⇒ ∅

As for other numerical expressions, the most common rules and most uniformly applied are the following:

1. Replacing a word with a figure:
   cinco días ⇒ 5 días

2. Rounding of big numbers:
   más de 540.000 personas ⇒ medio millón de personas

3. Rounding by elimination of decimal points:
   Cerca de 1,9 millones de casas ⇒ 2 millones de casas

4. Simplification of noun phrases containing two numerals in plural and the preposition *of* by eliminating the first numeral:
   cientos de miles de personas ⇒ miles de personas

5. Substitution of words denoting a certain number of years (such as *decade* or *centenary*) by the corresponding number:
   IV centenario de su nacimiento ⇒ *400 años de su nacimiento*

6. The thousands and millions in big numbers are expressed by means of a word, rather than a figure:
   17.000 *casas* ⇒ 17 *mil casas*

We are currently working on implementing a numerical expression simplification module based on rounding and rewording rules derived from our corpus and previous study in the field (Bautista et al., 2011).

4.5 Named Entities

As with numerical expressions, the majority of named entities are eliminated as a result of summarization. Only those names that are relative to the theme of the text in question and which tend to appear throughout the article are kept. In the case of these examples, we have observed the following operations: abbreviation; disabbreviation; using full name instead of the surname alone, customary in newspaper articles; expanding the noun phrase [ArtDef + NCom][3] with the name of the referent; replacing the noun phrase [ArtDef + NCom] with the name of the referent; inversion of the constituents in the structures where a professional title is followed by the name of its holder in apposition; and a handful of other, less frequent changes. Table 3 summarizes the most common operations and illustrates them with examples from the corpus. As can be observed, some NE are written as acronyms while others are disabbreviated. It would be interesting to analyse in the future whether the length and the relative frequency of the words that make up these expressions are a factor, or these are simply examples of arbitrary choices made by human editors lacking more specific guidelines.

While to decide how to deal with names of organisations that may possibly be abbreviated we would need a larger corpus more saturated in these examples, there are a number of rules ready to be implemented. Such is the case of personal names, where almost 90% of the names appearing in simplified texts contain both name and surname as opposed to first name alone. The same is true of the order of name and title, where in 100% of such examples the name is preferred in the initial position. As for expanding the named entity with a common noun (*the painter Pablo Picasso*), we have recorded this operation in only 15% of the personal names used in S texts. We do, however, notice a pattern — this kind of operation is applied at the first mention of the name, where the common noun acts as an additional defining element. It is an interesting phenomenon to be further researched.

4.6 Other simplification tendencies

Human editors have opted for a number of other simplification solutions which are either difficult or impossible to implement computationally. The elimination of nominalisations is an example of the former. Whereas common in the journalistic genre, human simplifications show a very strong tendency towards substituting the combination of the support verb and a deverbal noun with the corresponding verb alone, as in the example:

1. La financiación ha sido realizada por la Generalitat Valenciana.
   The funding has been provided by the Valencian Government.

2. La Generalitat Valenciana ha financiado la investigación.
   The Valencian Government has financed the research.

The expression *realizar una financiación* (*provide funding*) from the original sentence (1) has been substituted by the verb *financiar* (*to fund*) in the simplified version (2). Twenty other instances of this kind of operation have been recorded, thus making it an issue to be readdressed in the future.

What is also to be addressed is the treatment of set expressions such as idioms and collocations. Although not excessively abundant in the current version of our corpus, we hypothesise that the simplification of such expressions could considerably enhance the readability of the text and the research of the issue could, therefore, prove beneficial, provided

[3] NCom: common noun

Original	**Simple**	**Operation Type**
Comité Español de Representates de Personas con Discapacidad	CERMI	abbreviation
el PSOE	el Partido Socialista Obrero Español	disabbreviation
Gonzales-Sinde	Angeles Gonzales-Sinde	full name
el artista	el artista Pablo Picasso	NCom+NE
la ciudad andaluza	Granada	NCom ⇒ NE
La ministra de Defensa, Carme Chacón	Carme Chacón, ministra de Defensa	NCom,NE ⇒ NE,NCom

Table 3: Named Entities Substitution Examples

the availability of the necessary resources for Spanish.

On the other hand, an example of common human simplification tactics which is out of reach for a computational system is rewording of metaphorically used expressions. Thus, *un gigante de la escena* (*a giant on stage*) is changed into *un actor extraordinario* (*an extraordinary actor*). Such examples point out to the limitations automatic simplification systems are bound to possess.

5 Conclusions and future work

In the present paper we have concentrated on the analysis of lexical changes observed in a parallel corpus of original and simplified texts in Spanish. We have categorized all the operations into substitution of lexical units; insertion of definitions of difficult terms and concepts; simplification of numerical expressions; simplification of named entities; and different cases of rewording. Analysis suggests that frequency in combination with word length is the necessary combination of factors to consider when deciding on the choice among a set of synonyms to replace a difficult input word. On the other hand, a high number of polysemic input words underline the importance of including word sense disambiguation as part of the lexical substitution module.

Based on the available data, we have so far derived a set of rules concerning reporting verbs, adjectives of nationality, numerical expressions and named entities, all of which are to be further developed and implemented in the future. Numerical expressions in particular are given an important place in our system and more in-depth analysis is being carried out. We are working on rounding of big numbers and the use of modifiers in the simplification of these expressions. A number of issues are still to be tackled, such as elimination of nominalisation and simplification of multi-word expressions. The ultimate goal is to implement the lexical module as part of a larger architecture of the system for automatic text simplification for Spanish.

Acknowledgements

The present study is part of a Spanish research project entitled Simplext: An automatic system for text simplification (`http://www.simplext.es`). Simplext is led by Technosite and partially funded by the Ministry of Industry, Tourism and Trade of the Government of Spain, through the National Plan for Scientific Research, Development and Technological Innovation (I+D+i), within the strategic Action of Telecommunications and Information Society (Avanza Competitiveness, with the file number TSI-020302-2010-84). We are 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

A. Anula. 2008. Lecturas adaptadas a la enseñanza del español como l2: variables lingüísticas para la determinación del nivel de legibilidad. In *La evaluación en el aprendizaje y la enseñanza del español como LE/L2*.

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

S. Bautista, 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*.

S. Bautista, R. Hervás, P. Gervás, R. Power, and S. Williams. 2011. How to make numerical information accessible: Experimental identification of simplification strategies. In *Conference on Human-Computer Interaction*, Lisbon, Portugal.

O. Biran, S. Brody, and N. Elhadad. 2011. Putting it simply: a context-aware approach to lexical simplification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 496–501, Portland, Oregon, USA. Association for Computational Linguistics.

S. Bott and H. Saggion. 2011. An unsupervised alignment algorithm for text simplification corpus construction. In *ACL Workshop on Monolingual Text-to-Text Generation*, Portland, USA, June 2011. ACL, ACL.

Stefan Bott and H. Saggion. 2012a. Text simplification tools for spanish. In *Proceedings of Language Resources and Evaluation Conference, 2012*.

Stefan Bott and Horacio Saggion. 2012b. A hybrid system for spanish text simplification. In *Third Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, Montreal, Canada.

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

J. Carroll, 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.

H. M. Caseli, 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)*.

R. Chandrasekar, D. Doran, and B. Srinivas. 1996. Motivations and methods for text simplification. In *COLING*, pages 1041–1044.

H. Cunningham, 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*.

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

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

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

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

J. Medero and M. Ostendorf. 2011. Identifying targets for syntactic simplification.

Ll. Padró, 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.

S. E. Petersen and M. Ostendorf. 2007. Text simplification for language learners: a corpus analysis. In *Proc. of Workshop on Speech and Language Technology for Education*.

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

Lucia Specia. 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.