Lang Resources & Evaluation (2014) 48:93–120
DOI 10.1007/s10579-014-9265-4

SI: RESOURCES FOR LANGUAGE LEARNING

Text simplification resources for Spanish

Stefan Bott · Horacio Saggion

Received: 8 October 2011 / Accepted: 8 February 2014 / Published online: 2 March 2014

Abstract In this paper we present the development of a text simplification system for Spanish. Text simplification is the adaptation of a text for the special needs of certain groups of readers, such as language learners, people with cognitive difficulties, and elderly people, among others. There is a clear need for simplified texts, but manual production and adaptation of existing text is labour-intensive and costly. Automatic simplification is a field which attracts growing attention in Natural Language Processing, but, to the best of our knowledge, there are no existing simplification tools for Spanish. We present a corpus study which aims to identify the operations a text simplification system needs to carry out in order to produce an output similar to what human editors produce when they simplify news texts. We also present a first prototype for automatic simplification, which shows that the most important simplification operations can be successfully treated.

Keywords Text simplification · Aligned monolingual corpora · Simplification operations

1 Introduction

Reading not only provides us with the grateful experience of acquiring knowledge and expanding our horizons, but also helps us keep informed. However, text reading and understanding cannot be taken for granted. Some people can read different types of documents, from detailed technical literature to intricate novels, while others may find it difficult to read newspaper stories. The difficulty of a text depends on different factors, such as the reader herself and her personal experience, the

S. Bott · H. Saggion (✉)
Universitat Pompeu Fabra, C/Tanger 122, 08018 Barcelona, Spain
e-mail: horacio.saggion@upf.edu

social and historic context and the linguistic complexity of the text. Some readers may not be able to read and understand a text without it being adapted to their particular personal characteristics. In second language acquisition the preparation or adaptation of textual material which matches the abilities of the reader requires a heavy load of human resources. Adapted material can be prepared in various ways: (1) the material may consist of a completely new text specially designed for the target audience; (2) existing texts can be augmented or explained so that their content may become more comprehensible for the target audience; or (3) existing input material prepared for certain audience can be simplified in content and form to make it easy-to-read and understandable to the current target audience. It is this last type of adaptation which we address in this paper. Various organizations are dedicated to the production of textual material adapted to the needs of specific user communities. One methodology commonly used by these organizations to produce such adaptable textual material is the "easy-to-read" method (Petz and Tronbacke 2008). However, its dependence on human expertise and resources seriously limits the number of simplified content which can be produced based on existing texts. Automatic text simplification is a technology for the production of simplified texts, aiming at reducing, at least in part, the efforts required by manual simplification. Text simplification can reduce the syntactic, semantic, and lexical complexity of a given text by producing a quasi-paraphrase which will contain simple sentences, expressed in common vocabulary, depending on the needs of the intended audience or subjects. There are many users who could benefit from text simplification: people with low literacy level, deaf people, people with aphasia, immigrants who do not master the language of the receiving country, elderly people, second language learners, etc.

Most people with a minor cognitive disability are able to read and understand a text, and even people with major cognitive disabilities, although they may be unable to read, still enjoy someone reading a simple text to them. The UN Declaration of the Rights of the Persons with Disabilities emphasizes in its rule number five that content providers should make their content initially accessible or adapted to people with disabilities.[1] It is therefore important that governments and organizations alike implement simplification services.

Our research is concerned with the development of a text simplification system for Spanish. The simplification solution we are developing has an educational and social function, since its target users are people with cognitive disabilities. Our work is carried out within the Simplext project (Saggion et al. 2011) and our direct target group are students with Down Syndrome undertaking training at university programs designed for insertion in the marketplace. The initiative is lead by Prodis,[2] a foundation dedicated to the inclusion of disabled persons in education, work, and social areas.

Text simplification has been studied for some years in computational linguistics, with research undertaken for English, Portuguese, Japanese, French, Italian, Danish

[1] http://www.un.org/disabilities.

[2] http://www.fundacionprodis.org/.

and Basque but, to the best of our knowledge, there is no research in simplification for Spanish.

Although early work in simplification was undertaken without particular attention to the receiver of the simplified text (Chandrasekar et al. 1996), current work is generally targeted to specific user groups: people with aphasia (Carroll et al. 1998), people with low literacy level (Aluísio et al. 2008), language learners (Petersen and Ostendorf 2007), and deaf people (Inui et al. 2003). Dyslexic readers may only require simplifications at the lexical level (Hyönä and Olson 1995); on the other hand, it has been pointed out that aphasic readers find it hard to process sentences with passive voice or with pronouns (Devlin and Tait 1998). People with cognitive disabilities may benefit from simplification of the vocabulary as well as reduction of the syntactic complexity.

In this paper we present our work on the analysis of a dataset of original documents and their simplifications in order to identify, quantify and qualify text simplification operations. As it will be shown, the operations we have identified are in correspondence with traditional manual simplification methods. Our study identifies the operations that are most frequent and, among these, the ones that can be implemented in an automatic procedure. We demonstrate by means of an implemented prototype that the identified operations can be simulated as dependency tree transformations. The rest of the paper is organized in the following way: In the next section, we briefly describe the background of our work. In Sect. 3 we overview the method used in our research to produce manual simplifications, and in Sect. 4 we provide a detailed account of research in text simplification. In Sect. 5 we describe the dataset of simplifications we are using in this work, and then in Sect. 6 we show the analysis of the corpus. We qualify as well as quantify the operations and suggest ways for their computational treatment. In Sect. 7 we provide an implementation of the most relevant operations in the form of a prototype and discuss our approach in relation to the manual method and related work. We conclude the paper with our findings and avenues for further research.

2 Text adaptations

From a foundational and methodological point of view, there are various initiatives that promote accessible texts. An early proposal is Basic English, a language of reduced vocabulary of just over 800 word forms and a restricted number of grammatical rules. It was conceived as a tool for international communication or a kind of interlingua (Ogden 1937), and it was promoted after the Second World War as a means of international communication. Other initiatives are Plain English (Brown 1995), for English in the United States and in the United Kingdom, and the Rational French, a French controlled language to make technical documentation more accessible in the context of the aerospace industry (Barthe et al. 1999).

In Europe, there are associations dedicated to the adaptation of text materials (books, leaflets, laws, official documents, etc.) for people with disabilities or low literacy levels, examples of which are the Easy-to-Read Network in the

Scandinavian countries, the Asociación Lectura Fácil[3] in Spain, and the Centrum för Lättläst in Sweden.[4]

These associations usually provide guidance or recommendation about how to prepare/adapt textual material. Recommendations will be of the kind:

- Simple and direct language
- One idea per sentence
- Avoid jargon and technical terms
- Avoid abbreviations
- Structure text in a clear and coherent way
- One word per concept
- Personalization
- Active voice

These recommendations, although understandable, are sometimes difficult to operationalize (for both humans and machines) and sometimes even impossible to follow, especially in the case of adapting an existing piece of text.

In addition to books and specially prepared material, there is a plethora of simplified material on the Web. The Swedish "easy-to-read" newspaper "8 Sidor"[5] is published by the Centrum för Lättläst to allow people access to "easy news". Other examples of similarly oriented online newspapers and magazines are the Norwegian Klar Tale,[6] the Belgium l'Essentiel,[7] the Flamish Wablie,[8] the Danish Radio Ligetil,[9] the Italian Due Parole,[10] and the Finnish Selo-Uutiset.[11] For Spanish, the Noticias Fácil website[12] provides easy-to-read news for people with disabilities. The Literacyworks web site[13] offers CNN news stories in original and abridged (or simplified) formats, which can be used as learning resources for adults with poor reading skills. Around 104 original and abridged parallel news stories were used for investigating automatic text simplification operations for second language learning, such as sentence splitting and sentence elimination (Petersen and Ostendorf 2007). At the European level, the Inclusion Europe Web site[14] provides good examples of how full text simplifications and simplified summaries in various European languages can provide improved access to relevant information. The Simple English Wikipedia[15] provides encyclopedic content

[3] http://www.lecturafacil.net/.

[4] http://www.lattlast.se/.

[5] http://8sidor.lattlast.se.

[6] http://www.klartale.no.

[7] http://cours.funoc.be/essentiel/.

[8] http://www.wablieft.be.

[9] http://www.dr.dk/Nyheder/Ligetil/Presse/Artikler/om.htm.

[10] http://www.dueparole.it.

[11] http://papunet.net/selko.

[12] http://www.noticiasfacil.es.

[13] http://www.literacyworks.org/learningresources/.

[14] http://www.inclusion-europe.org.

[15] http://simple.wikipedia.org.

which is more accessible than plain Wikipedia articles because of the use of simple language and simple grammatical structures.

The number of web sites containing manually simplified material pointed out above clearly indicates a need for simplified texts. However, manual simplification of written documents is very expensive and manual methods will be not cost-effective, especially if we consider simplification of news which are constantly produced.

3 Manual simplification methodology

We are developing a corpus of manually simplified informative articles (newspaper stories), which serves as the basis for the development of language resources and algorithms for the study of Spanish text simplification, and the development of a practical simplification solution.

The manual simplification methodology adopted in our work in order to produce simpler texts is based on an easy-to-read method adapted for people with cognitive disabilities (Anula 2007, 2008), which is our target group. The application of this methodology has been proven to contribute to the reduction of complexity in written language and to make texts easier to read. It considers two variables for measuring text complexity: vocabulary complexity and syntactic complexity. The objective of the method is to reduce the values of these variables in the simplified text. Specific methods are designed to measure the two variables. We cannot give a full account of the manual simplification process here, which contains over 30 simplification recommendations and rules, since that would be beyond the scope of this paper, but we will comment on some of the simplification operations that have been proposed.

Where the lexicon is concerned, the method proposes to look at word frequency and to replace uncommon words by their more common synonyms, looking up the frequency in a reference corpus such as the Real Academia Española corpus.[16] It also addresses word ambiguity and attempts to replace word forms which are ambiguous by their less ambiguous synonyms, if possible. This adaptation is not trivial from the natural language processing point of view since: (1) many words are polysemous, thus complicating the process of string replacement; and (2) the context of the word to be replaced could require some linguistic manipulation for the final text to be correct (e.g. consider the replacement of the word *problema* (*problem*), which is masculine, with the word *complicación* (*complication*) which is feminine). A simple text should generally use the same word to express the same concept; therefore, if possible, a concept will be referred to by the same word in the simplified text. The effect of this simplification is the reduced cognitive effort associated with meaning recovery. Replacing the same concept by a unique linguistic expression is also complicated for natural language processing tools since it requires appropriate treatment of anaphoric phenomena.

Note that the notion of word simplicity was loosely defined in the English lexical simplification task in SemEval 2012 (Specia et al. 2012) as words which can be

[16] Corpus de referencia del español actual, http://www.rae.es.

understood by a wide range of people, including those with low literacy levels or some cognitive disability, children, and non-native speakers of English.

Where the sentences and discourse segments are concerned, simple texts should have shorter sentences. DuBay (2004) reports findings indicating that in English "very easy" texts have an average sentence length of 8 words, "easy" texts an average sentence length of 11, and "fairly easy" texts an average sentence length of 14. For Spanish, we found that in our parallel corpus the average number of words per sentence in simplified texts is 12.44 words and in the original 34.64. The minimum number of words is nearly the same in both simple and original corpus texts, 4 and 5 words respectively, but the maximum number of words we observed is very different: in the simplified texts the longest sentence had 24 words while in the original the maximum length was 88 (a sentence containing a long enumeration of names of people and organizations). This implies that long sentences in a text should be split in order to make them simple. Coordinations and subordinations should be simplified transforming them into several independent sentences when possible.

Some functional words are preferred over others; for example instead of using *sin embargo* (*however*), a more common form such as *pero* (*but*) should be used. Particular attention has to be paid to these replacements since, for example, the expression *pero sin embargo* (*but however*) could not be blindly replaced by *pero pero* (*but but*). The context is also very important for lexical simplification, since it determines in which sense a content word must be interpreted (Bott et al. 2012). For example the Spanish word *hogar* usually means *home* and is synonymous with the more frequent word *casa* (*house*), but depending on the context it can also mean fireplace (e.g. *el fuego arde en su hogar/the fire burns in the fireplace*).

Content reduction is another method of reducing the complexity of a sentence, but this should be applied to elements which are unnecessary for the understanding of the whole text. Text summarization techniques could be of value in order to implement such operations (Saggion 2008). A simpler text could also contain extra information not present in the original, if this makes the text more understandable. A typical situation would be the inclusion of a definition of a term that the reader probably does not know.

In Sect. 6 we quantify and qualify the simplification operations in order to understand which of them should and could be implemented in an automatic system. We will show that similar operations to those presented here emerge from the data.

4 Related work

From the Natural Language Processing point of view, Text Simplification has been studied with the following objectives:

- to allow people with low literacy level access to information, which in turns facilitates social inclusion (Watanabe et al. 2009; Aluísio et al. 2008);

- to make news paper articles accessible for people with a reduced intellectual ability (Carroll et al. 1998; Max 2006) or for people who need assisted reading (Inui et al. 2003);
- to facilitate access to textual information to foreign readers (Crossley and Mcnamara 2008);
- to allow access to texts of high complexity such as patents, regulations, laws, etc., to people unfamiliar with the intricacies of these texts (Bouayad-Agha et al. 2009);
- to reduce the complexity of natural language texts in order to facilitate tasks such as syntactic and semantic analysis or machine translation (Klebanov et al. 2004; Ong et al. 2008).

Early attempts of automatic text simplification used rule-based methods, where rules were designed following linguistic intuitions (Chandrasekar et al. 1996). Steps in the process included linguistic text analysis (including parsing), pattern matching, and transformation steps. Other computational models of text simplification included the processes of analysis, transformation, and phrase re-generation, also using rule-based techniques (Siddharthan 2002). In the PSET project (Carroll et al. 1998; Canning et al. 2000), a news simplification system for aphasic readers is proposed, where particular attention is paid to linguistic phenomena such as passive constructions and co-references, which are difficult to deal with by people with disabilities. The European PATExpert project (Mille and Wanner 2008; Bouayad-Agha et al. 2009) has developed a simplification technology to address the problem of readability of patent claims. This approach makes use of segmentation of long patent claims into short units, establishment of co-reference links between units, and reconstruction of each unit using text generation technology. The PorSimples project (Aluísio et al. 2008; Gasperin et al. 2010) has looked into simplification of the Portuguese language. The methodology consisted of the creation of a corpus of simplification at two different levels, and the use of the corpus to train a decision procedure for simplification based on linguistic features. Simplification procedures for different linguistic phenomena are implemented and applied in cascade to the input text. Decisions about whether to simplify a text or sentence have been studied following rule-based paradigms or trainable systems (Petersen and Ostendorf 2007), where a corpus of texts and their simplifications becomes necessary. Max (2006) proposes the integration of automatic simplification into a text authoring system, a step which could give automatic simplification more practical relevance in the creation of accessible material. More recently Siddharthan (2011) compared a rule-based simplification system with a simplification system based on a general purpose generator. Aranzabe et al. (2012) describe a text simplification system for Basque, a language which is especially challenging because it has limited resources and is difficult to treat due to typological reasons, being an agglutinative language. Seretan (2012) presents a method for semi-manual acquisition of simplification rules for French with the aim of implementing an automatic system. Daelemans et al. (2004) compare two approaches for simplifying sentences with the purpose of using them in a subtitling application. One approach learns word deletion and word replacement simplification operations using a parallel corpus. The second approach, which

outperforms the first one, is rule-based and replaces full noun phases with their heads, deletes adjectives and adverbs in specific contexts, and drops conjunctions in initial sentence positions.

Related to the problem of automatic simplification is the problem of measuring text readability with the objective of developing metrics able to associate a readability score to the texts. For the English language, there is the well-known test developed by Flesch (1948), which combines the mean length of sentences with the average number of syllables per word. A recent proposal looked into semantic content or entity density (Feng et al. 2009, 2010) as a feature that, when combined with syntactic features, improves classification according to readability scales. The Coh-Metrix tool (Graesser et al. 2004), which computes over 60 different text indices, can be used to measure text complexity. There are also studies in text readability for languages other than English; for example, for Spanish, Rodríguez Diéguez et al. (1992) introduce 12 variables such as common noun and proper noun distribution, punctuation distribution, etc., to produce a combined readability score. For French, Tanguy and Tulechki (2009) proposed a set of automatically derived features to measure complexity of text and sentences. Dell'Orletta et al. (2011) created a text and sentence classification system for Italian. They developed this tool specifically for the use in the context of automatic text simplification, and for this task they stress the importance of assessing the relative simplicity of individual sentences. An interesting finding they made concerns syntactic and morpho-syntactic features, like lexical density and syntactic embedding depth. For the classification of texts (into simple and non-simple), the inclusion of these features in the classifier improves the result only slightly, but when it comes to the classification of individual sentences, the same features are much more important, and including them considerably improves the classifier.[17] Pitler and Nenkova (2008) studied the problem of text quality prediction which includes text readability as one of several factors. Their framework identified and assessed the correlation between a series of textual, syntactic, semantic, and discursive features with text quality, finding that the distribution of discourse relations in sentences as well as vocabulary and text length are good predictors of text quality.

Very little work has been done building and using parallel corpora for text simplification until recently. Some resources are available for the English language, such as parallel corpora created or studied in various projects (Barzilay and Elhadad 2003; Feng et al. 2009; Petersen and Ostendorf 2007; Bouayad-Agha et al. 2009). In Specia (2010), a manually built corpus of complex and simplified Portuguese texts for poor literacy readers is used and techniques from phrase-based statistical machine translation were applied to learn how to 'translate' complex into simplified texts. The system focuses on lexical simplification and simple reordering operations (of mostly adverbs), since it does not use any syntax. In Jing (2002), a summarization system that includes sentence reduction (which for our purposes can be seen as a form of simplification) is developed, where one of the key features is that it utilizes a corpus consisting of original sentences and their corresponding

[17] They report an improvement from 61.6 to 78.2 % of accuracy when including syntax and morpho-syntax in addition to basic counts and lexical information.

reduced forms written by humans for training and testing purposes. Using this corpus in conjunction with syntactic parsing and grammar checking, they identified six operations that can be used alone or together to reduce extracted sentences. Zhu et al. (2010) propose a system for English text simplification based on the statistical machine translation framework. However, they use syntactic information to improve the generated rules. The parallel corpus used in this paper is automatically generated using articles from Simple English Wikipedia and their corresponding English Wikipedia (complex) versions. Coster and Kauchak (2011) also used the Wikipedia resource and applied a machine translation framework to implement lexical substitution. Woodsend and Lapata (2011) use quasi-synchronous grammars as a more sophisticated formalism and integer programming in order to learn to translate from English to Simple English. This system can handle sentence splitting operations, and the authors use both automatic and human evaluation and show an improvement over the results of Zhu et al. (2010) on the same data set, but they have to admit that learning from parallel bi-text is not as efficient as learning from revision histories of the Wiki-pages.

One of the main operations in text simplification systems is that of replacing words by simpler synonyms. Lexical simplification has been usually addressed by simple lookup in thesauri or databases with frequency information. For English, the standard resource was developed using WordNet, and psycholinguistic and frequency information (Devlin and Tait 1998). In such approaches, the context of the complex target word is disregarded. An exception is the work by De Belder et al. (2010) in which word sense disambiguation is performed to choose among a set of possible simplifications.

5 Data and data preparation

As we mentioned in Sect. 3, we are preparing a parallel corpus of 200 news paper articles with their manually simplified counterparts, covering the topic domains of national news, international news, society and culture. The simplified versions are provided by trained experts from the DILES research group from the Universidad Autónoma de Madrid, and are based on the guidelines by Anula (2011). Each text was simplified by one editor and revised by several members of DILES. We have, however, no direct insight into the editing process itself in the form of editing histories. The size of the corpus is not big enough to make pure machine learning techniques a promising option for text simplification as a global problem. There are no other large parallel text resources for simplified Spanish which could serve as an empirical basis for data-driven methods, like the Simple English Wikipedia, which has recently received attention as a resource for text simplification approaches (Coster and Kauchak 2011; Zhu et al. 2010).

The texts are processed using parts-of-speech tagging, named entity recognition and parsing (Padró et al. 2010), in order to create an automatically annotated version of the corpus. We developed an automatic alignment tool, which aligns the texts of the corpus on the sentence level (Bott and Saggion 2011). The automatic alignment is then hand-corrected with a graphical editing tool which is based on the

GATE framework (Maynard et al. 2002). Sentence alignment is crucial for corpus studies and also for possible machine learning experiments, which might be conducted on specific sub-problems of text simplification.

For the current version of the simplification prototype we use a dependency parser (Bohnet 2009) and the tree-transduction tool MATE (Bohnet et al. 2000). MATE is a tool which was created with the mapping between different layers of linguistic representation in mind, and it is especially useful for text generation. However, in our context, we use MATE as a tool that maps syntactic dependency structures which we detect as requiring simplification onto simplified versions of these structures. MATE allows the creation of syntactic rules which manipulate syntactic structure of an arbitrary depth. We will show some examples of this in Sect. 7.

At present, we only use information which comes directly from the dependency parser in our simplification rules. We are, however, working on the integration of information stemming from additional sources. We especially encountered the need to have access to named entity categories and information about nominal co-reference. Named Entity Information is important especially for the recovery and copying of subjects in cases where a subordinate clause is turned into a separate sentence and a subject for this sentence has to be inserted. Without the information about named entities it is often very hard to delimit the extend of a multi-word NP (e.g. *Red Cross*) which has to be used as the subject. Nominal co-reference is important to change a pronoun into a full noun, an operation which we found in the human simplified texts and which tries to make all nominal references as explicit as possible.

6 Simplification edit operations

In order to evaluate which operations we have to cover in the implementation of our simplification tool, we examined the corpus and developed an annotation scheme for the simplification edit operations we could find. The corpus is still under development and when we carried out the corpus study we only had a sample of 145 sentences available. Nevertheless, we believe that this corpus fragment is representative of the whole corpus: it is composed of short news texts and, more importantly, manual simplifications were carried out on the basis of the same simplification guidelines and by the same team of expert editors.

We decided to annotate the changes found in the bi-text in two different dimensions. The first dimension represents the main classes of simplification operations—for example, whether material has been deleted or inserted, or if we could observe changes in the syntactic realization. Since many of these operations can affect linguistic units of different types, we used a second dimension of annotation, which represents the type of unit affected. As for the first dimension, we could find eight major operation types:

- change
- delete

- insert
- split
- proximization
- reorder
- select
- join

There have been earlier classifications of simplification operations, but we found that, for our needs, these taxonomies were not detailed enough. Chandrasekar et al. (1996) concentrate on sentence splitting. De Belder et al. (2010) concentrate, in turn, on lexical simplification. Zhu et al. (2010) list 4 simplification operations: deletions (also called *dropping*), lexical change (*substitution*), reordering, and splitting. Coster and Kauchak (2011) have a similar list, but they also include *insertions*. We tried not to rely on earlier classifications and create our own taxonomy from manual inspection of the data. Since our automatic simplification approach necessarily involves manual creation of simplification rules, we were, in addition, interested in more detailed subclasses of these top-level operations. As already said, the major operation types listed here represent only one dimension of annotation and are independent from the linguistic level at which they apply. *Change, delete* and *insert* operations can apply at the word level or affect larger syntactic units. A more fine-grained classification will be discussed below.

The annotation scheme is also independent from prescriptive or suggestive simplification guidelines for human editors that carry out text simplification, since the guidelines do not normally try to classify their recommendations in categories like the ones mentioned here. There are several reasons why we did try to use human-oriented guidelines for the creation of our annotation set. Firstly, we were interested in what human editors actually do when they simplify texts, not in what they are expected or suggested to do. Most simplification guidelines are vague, underspecified and ambiguous: they often tell the editor which operations to avoid and sometimes expect her to find alternative solutions which are not further specified. If alternatives are listed for a target construction, there are often several possibilities from which the editor has to choose according to her criterion. Finally, simplification guidelines often expect the editor to make inferences on the basis of contextual and world knowledge, far beyond the inferences that a computer can be expected to carry out at present.

The raw frequencies of these operations, given in Table 1, already give us a rough idea of the importance of each editing operation. *Change* operations are by far the most frequent ones. These include, above all, lexical changes, but also syntactic changes of various types and changes of the verb form, such as voice. *Delete* operations cut out syntactic units which are considered to convey little additional information, such as adverbial phrases and adjectives. *Insert* and *split* operations are somewhat less frequent, but their effect on text complexity is usually quite marked. *Insert* operations recover implicit information from the context and make it implicit in a target sentence. *Splits* are a per excellence syntactic simplification operation: they convert embedded or coordinated clauses in independent sentences and reduce both sentence length and, in particular, syntactic embedding depth.

Table 1 Frequencies of different editing operations

Operation	%
Change	39.02
Delete	24.80
Insert	12.60
Split	12.20
Proximization	6.91
Reorder	2.85
Select	0.81
Join	0.81

The different edit operations interact in complex ways. We found an average of slightly more than four edit operations per source sentence. The examples quoted below reflect this rich interaction and it is, actually, difficult to find examples which illustrate only one edit operation in isolation.

As already mentioned, the mayor simplification operations are orthogonal to the levels of linguistic representation and do not represent the specific linguistic phenomena that simplification guidelines typically list. The goal of the corpus study is to determine which types of computational operations a simplification system has to be able to carry out, how much impact these operations have in terms of frequency and their influence on text complexity, and how difficult they are to implement computationally. We therefore tried to further specify the list given above, according to the linguistic level and the grammatical constructions they affect. This is not always a trivial task, since human editors tend to produce quite complex rewritings, instead of only performing clear-cut editing operations.

Table 2 lists the most frequent operation subtypes. This table introduces the second dimension to the annotation: While Table 1 only lists the type of operation, Table 2 specifies the linguistic target category or grammatical construction type (with the exception of *proximizations*, which could not be associated to a given category and shall be described below). For example, the operation *change: full clause →NP* affects a full clause and compresses it to a noun phrase (for the creation of a headline), while the category *change: syntax* only affects a part of a sentence. We left aside those cases which were very idiosyncratic, involved free rewording or were otherwise very hard to classify. Those cases were very frequent (43 % of all observed operations), a finding which reflects the fact that human editors tend to produce very free paraphrases, which are hard to capture in computational terms. In the remainder of this section, we will discuss the most important categories listed and draw some conclusions on in how far they can be implemented in an automated simplification system.

The most frequent of these cases were lexical substitutions, which represented 17.48 % of all edits. Here "difficult words" are replaced by their simpler counterparts. Words that are considered difficult include very infrequent words, very long words and foreign words, as well as technical terms. Example (1) shows an example of lexical substitution (and a *coordination split* which will be discussed

Table 2 Frequencies of specific editing operations

Operation	%
change: lexical	17.48
delete: adverbial or adjectival	7.32
proximization	6.91
delete: clausal	4.07
insert: unrestricted	3.66
change: syntax	2.85
change: voice	2.44
split: coordination	2.44
split: relative clause	2.03
insert: missing main verb	1.63
split: participle and gerundive construction	1.22
change: pronoun → full noun	0.81
change: full clause → NP	0.81
reorder: direct speech	0.81
insert: missing inflected verb	0.41
change: numbers	0.41
split: subordinate clause	0.41

below). Here the words *fauna* and *botánica* (*botanics*), which are more formal, are substituted by the more commonplace words *animales* (*animals*) and *plantas* (*plants*) respectively. The need for lexical simplification is also stressed in the easy-to-read methodology, described in Sects. 2 and 3. This type of operation is easy to implement, since it only involves string substitutions or substitutions of very shallow syntactic tree fragments.

(1) orig: La muestra ofrece al público la oportunidad de acercarse a la fauna, la botánica y la cultura de esta inmensa región selvática americana (…).
"The exhibition offers the public the opportunity to get close to the fauna, the botanics and the culture of this immense American jungle region."

simp: La exposición nos muestra la cultura de esta gran selva americana. También nos muestra sus animales y plantas (…).
"The exhibition shows us the culture of this great American jungle. It also shows us its animals and plants."

Change operations can also operate on the syntactic level. In (2), an appositive construction is turned into copulative construction and presents the headline type NP as a full sentence with a main verb. Even if this is not the case in this particular example, appositive constructions often result in sentence splitting, since the

apposition contains additional information about the nominal referent to which it is syntactically attached.[18]

(2) orig: "Escuela Segura, un compromiso municipal con la protección integral de los escolares."
"Safe School: a municipal promise for the full protection of school kids."
simp: Escuela Segura es un programa municipal para la protección de los escolares.
"Safe School is a municipal promise for the full protection of school kids."

In (3) the *change* affects the voice of the main verb, which is converted from active to passive (which is combined with relative clause split and the deletion of the first clause).

(3) orig: Se trata de un proyecto (…) que coordina el trabajo (…) de las delegaciones municipales de Educación y Seguridad (…)
"It consists of a project that coordinates the work of the city's education and security delegations."
simp: El proyecto está coordinado por las delegaciones municipales de Educación y Seguridad (…).
"The project is coordinated by the city's education and security delegations."

The subject of the active clause is not agentive, so an impersonal passive formulation is preferred. Note that the particular simplification step in this example goes against the general recommendation of the easy-to-read method to use only active voice (cf. Sect. 2). However, the original recommendation was formulated for English, which has a much wider use of passives than Spanish, and the decisive factor here seems to be to avoid constructions which are introduced by the frequent Spanish formulation *se trata de* (*it consists of*). Language specific traits often override more general concerns. Given a reliable syntactic representation, these changes are not difficult to carry out, but the cases which require simplification are hard to detect automatically. By no means do we want to change all verbs from active into passive in the same way as in (3). In order to spot such examples we would have to detect non-agentive subjects, a thing which is hard to do with current semantic role labeling systems. An additional problem in this example is that it includes an inference, even a defeasible one: in this case it is true that the project is coordinated by the delegations, but in order to support this inference, the context must be taken into account. For computational systems such inferences are practically impossible to perform.

[18] For example the sentence *Álex de la Iglesia, the director of the Academy, announced his resignation* contains the information that *Álex de la Iglesia is the director of the Academy*, which can be expressed in a separate sentence.

A further type of syntactic change expands pronouns into full noun phrases recovered from the context, making the sentence argument structure more explicit. This last type of operation requires reliable co-reference resolution and is thus computationally more challenging. Another, somewhat idiosyncratic, change operation is *numeric change*, which implies rounding of large numbers and even supplementing them with verbal expressions, like in *más de* (*more than*) in (4). Round numbers have been argued to be easier to remember and to calculate with (Krifka 2007). Power and Williams (2011) present a computational method of creating such rounded expressions in English. For Spanish, a numeric rounding module needs to be implemented.

(4) orig: (…) se han plantado 241 árboles y 4.377 arbustos.
"241 trees and 4,377 bushes have been planted."
simp: Hemos plantado 241 árboles y más de 4000 plantas.
"We planted 241 trees and more than 4,000 plants."

Another frequent group of edit operation is represented by deletions of adjectives, adverbs and adverbial phrases. In (5) the adjective *diversas* (*several*) is deleted, because it carries very few additional information. It is easy to carry out these deletion operations as pruning of a syntactic tree, but it is less easy to determine which adjective, adverbs or adjuncts are actually semantically light enough to justify such a deletion. Deletion operations nearly always imply a *content reduction*, which can affect content that is necessary for the understanding of a text as a whole. *Deletion* can also affect whole clauses. Again the decision to delete a clause is hard to make while the deletion itself is usually easy to perform.

(5) orig: Sanse coopera con diversas comunidades de Bolivia y Guatemala.
"Sanse cooperates with several communities in Bolivia and Guatemala."
simp: Sanse coopera con ___ comunidades de Bolivia y Guatemala.
"Sanse cooperates with communities in Bolivia and Guatemala."

Insertion operations may also affect different levels of linguistic representation. Unfortunately, we found that the most frequent among the insertion operations are *unrestricted insertions*. This label includes a mixed bag of insertion operations which were hard to construe as real editing operations on a source text and seemed to involve a good deal of inference made by the editor. Given this, it seems unlikely that a computational system can possibly perform them. (6) is an example of such unrestricted insertions. Here the editor inferred that the information expressed in a fragment expressed the office hours. While it is a feasible task to detect something like opening hours or telephone numbers, the list of unrestricted insertion operations is very long and not all the specific operation are very frequent. So the work of implementing specific dedicated modules does not seem promising.

(6) orig: De 9:30 a 14 horas y de 15 a 17:30 horas.
"From 9:30h to 14:00h and from 15:00h to 17:30h."
simp: El horario será de 9:30 a 14 horas y de 15 a 17:30 horas.
"The office hours will be from 9:30 to 14 and from 15 to 17:30."

While unrestricted insertions are problematic, there are insertion operations which are better defined and rescue otherwise ungrammatical sentences. In combination with split operations, the insertion of a main verb of an inflected verb form may be necessary. In (2), for example, the copula *es* is inserted in order to form a full sentence from the appositional NP. It is not difficult to create syntactic rules that can detect sentences without a main verb and insert a semantically light verb if necessary.

The family of *split* operations affects one of the key aspects of text complexity: the depth of clausal embedding. Editors often turn relative clauses or participle constructions (as in (7) and (8), respectively) into independent sentences. The complexity and length of sentences is reduced in this way. Splitting sentences also makes a sentence compliant with the easy-to-read recommendation to express only one idea per sentence.

(7) orig: 5.000 metros cuadrados (…) en los cuales se han plantado 241 árboles y 4.377 arbustos.
"5,000 square meters in which 241 trees and 4,377 bushes have been planted."
simp: El parque tiene más de 5000 metros cuadrados de zona verde. Hemos plantado 241 árboles y más de 4000 plantas.
"The park has more than 5,000 square meters of green areas. We have planted 241 trees and more than 4,000 plants."

(8) orig: 5.000 metros cuadrados situados entre las calles Doctor Fleming, Martín Chirino y Paseo de Europa en los cuales se han plantado 241 árboles y 4,377 arbustos.
"5,000 square meters situated between Doctor Fleming street, Martín Chirino street and Paseo de Europa, in which 241 trees and 4,377 bushes have been planted."
simp: El parque tiene más de 5000 metros cuadrados de zona verde. Está entre las calles Doctor Fleming, Martín Chirino y Paseo de Europa.
"The park has more than 5,000 square meters of green areas. Is is between Doctor Fleming street, Martín Chirino street and Paseo de Europa."

A further *split* operation affects clausal coordination. In this case, each coordinate is turned into a separate sentence, as exemplified by (1). Even if in this case the depth of embedding is not reduced, the average length of sentences becomes shorter. *Split* operations can be modeled as manipulations of the syntactic tree. A potential problem here is that these operations are very sensitive to parsing errors, especially in the case of coordination separation.

We found that there is a special operation, which we dubbed *proximization* (cf. Table 2), which is hardly comparable to the other editing operations. This operation type serves to make sentences psychologically closer to the reader. When the text is about an event in a certain city and this event is announced in the local newspaper, sometimes a locative phrase like *in our city* (*en nuestra ciudad*) may be inserted, or a third person verb form (*the interested person can*) is turned into second person (*you can*). These operations are often hard to predict when text simplification is taken to be a chain of editing operations. Example (4) shows, among other phenomena, an instance of proximization. An impersonal construction (*se ha plantado*) is substituted by a second person plural construction (*hemos plantado*). This operation is roughly equivalent to the easy-to-read recommendation to *personalize* texts for a specific reader group. Such substitutions require heavy inferencing, which is in many cases even defeasible (i.e. the inference that a certain city is the same as *our city*). So it seems very unlikely that this kind of operation can be automated.

Reorder operations change the order in which information is presented to the reader. A very typical case occurs with direct speech. Here the person speaking is commonly named before the quote in simplified texts, while in the original text often the person speaking is expressed after the quoted speech, in a clause separated by a comma. The example (9) shows this kind of operation, in addition to the expansion of a pronoun to a full NP and a lexical change.

(9) orig: "Con ellos ofrecemos una nueva posibilidad para (…) propiciar un envejecimiento activo y saludable así como una mejor calidad de vida", afirma Dolores de Diego(…).
" 'With these we offer the possibility to (…) achieve an active and healthy way of getting older, as well as a better quality of life', Dolores de Diego (…) points out."

simp: Dolores Diego (…) afirma: "Los aparatos propician un envejecimiento saludable y mejoran la calidad de vida de las personas mayores."
"Dolores Diego points out: 'The (exercise) machines offer a healthy way of getting older and improve the quality of life of elderly people.' "

The last two edit operations, which we have not described so far, are less common. *Select* operations may pick a NP out of a source sentence and use this NP as a headline, or single out a relevant piece of information and present it as a short sentence. In (10), this information is extracted from a longer section of quoted

speech. The relatively rare *join* operation combines two separate pieces of information into a single sentence, as in (11). In this example further information from the context was included for clarification.

(10) orig: "Hemos apostado (…) por el riego con agua reciclada (…), pues la primera razón de ser de un parque es su coherencia con la sostenibilidad", afirmó el alcalde.

" 'We opted (…) for irrigation with recycled water (…), as the first reason for a park is its consistency with sustainability,' said the mayor."

simp: El parque se riega con agua reciclada.

"The park is irrigated with recycled water."

(11) orig: Hasta el 22 de enero. De lunes a viernes.

"Until 22 January. Monday to Friday."

simp: Los exámenes se harán de lunes a viernes hasta el 22 de enero.

"The exams will be from Monday to Friday until 22 January."

We conclude this section with some reflections on the possibility to implement the operations we found in human simplifications in an automatic text simplification system. There are various factors which influence this: the computational feasibility itself, the frequency of operation type, the degree of variance within the operation type class, and the expected influence on the degree of reading ease. The edit operations described in this section fall on different points in a continuum which represents the feasibility of computational implementation. There are clearly some operations which we cannot hope to implement successfully, especially those which need inference from contextual information and reasoning on world knowledge. *Proximizations* and *unrestricted insertions* in particular seem very hard to carry out. On the other end of the spectrum we can find operations which can be processed easily. Lexical substitutions are the prototypical case of this class: they are very frequent and contribute, in addition, to a high degree to the reduction of reading difficulty. It has also been found that lexical simplification by itself can be helpful for users with some cognitive conditions, such as aphasic readers or people with dyslexia (Hyönä and Olson 1995). Other operations, especially the split operations, are slightly harder to model, but are quite frequent and reduce the structural complexity of the text. The split-operation we found have a high potential to simplify texts, since they both reduce the depth of the parse tree and the sentence length. In addition, although different split operations require dedicated rules, these operations have many similarities among them. Deletion operations also seem feasible, since they require techniques similar to those applied in text summarization and sentence compression. In the next section we describe a system, which addresses two of the simplification operations described: lexical changes and various types of split operations. We also take the corpus study presented here as a

starting point for the planning of future work, for example on context reduction (which covers various delete operations).

7 A first prototype

The corpus study described in Sect. 6 showed us that nearly all of the simplification operations can be modeled as operations on syntactic trees, even if the highly frequent lexical substitution operations represent only trivial changes on a syntactic tree, namely the substitution of a leaf. The different simplification operations do, however, need different types of information in order to be carried out. Lexical substitution needs a lexical resource, such as a dictionary of words to be substituted. Such a resource is being created on the basis of frequency counts extracted from a corpus.[19] Some insertion operations, as well as the substitution of pronouns, need information about the co-reference of nominal expressions. Finally, many of the operations need to take the context into account; for example, word meanings dependent on the textual domain and several other simplification functions have to recover antecedent NPs from the context.

In the creation of a prototype we concentrated on two operations: lexical substitution and syntactic split operations. The first is important because it is the single most frequent operation type. A satisfactory treatment of this simplification operation can significantly reduce the difficulty of reading a text. The substitution of single words is not very difficult, but sometimes a multi-word unit has to be substituted, and in a series of cases even the word order has to be changed. We therefore included multi-word substitution in the problem. In the prototype presented here, we mainly focus on this latter type of lexical simplification problem. Within the Simplext project we are also developing a statistical lexical simplification tool (Bott et al. 2012), which is motivated by the findings described in the last section. The second simplification operation we concentrated on is sentence splitting. Sentence splitting reduces the structural complexity of a text in terms of embedding depth and sentence length. We also considered that this operation is typical for the whole group of syntactic manipulations. The same problems encountered when converting one syntactic tree into several grammatically well formed trees can be expected to arise in other simplification operations: detecting a target configuration which needs simplification, manipulating the tree itself (modelled as tree transduction), and making sure that the output tree results in a grammatical sentence.

Lexical simplification often involves the substitution of a single word with a simpler synonym, but it also includes the simplification of function words and discourse markers, such as subordinate conjunctions, temporal conjunctions and adversative connectors, among others. Example (12) contains the complex adversative discourse marker *por el contrario* (*to the contrary*), which can be substituted with the much simpler and shorter word *pero* (*but*).

[19] Corpus de referencia del español actual, http://www.rae.es.

(12) Por el contrario, según la citada encuesta, el 72 % de los españoles hace una valoración positiva de la misión española en Bosnia.

"On the contrary, according to the cited poll, 72 % of the Spanish population have a positive opinion about the Spanish mission in Bosnia."

The syntactic dependency tree of (12), as produced by our parser, is given in Fig. 1. The tree is slightly simplified for expository purposes: the subtree corresponding to *according to the cited poll*, as well as the information about word order have been deleted.[20] The relevant parts of the tree have been highlighted.

We carry out the simplification of such structures in two steps: first the relevant lexicalized target structure is identified and marked for simplification, and in a second step the whole sub-tree is substituted, resulting in the structure shown in Fig. 2.

Sentence splitting can target various constructions which subordinate or coordinate clauses within a matrix clause. A typical case are relative clauses.

(13) Estos pisos son inventariados (…) y se ofrecen a los jóvenes solicitantes, a los que se acompaña en la visita.

"These flats have been registered and are being offered to young applicants, which will be accompanied during the visit."

Here the sentence can be simplified by using a separate sentence to express the fact that the applicants will be accompanied, as exemplified in (14):

(14) a. Estos pisos son inventariados (…) y se ofrecen a los jóvenes solicitantes.

"These flats have been registered and are being offered to young applicants."

b. A los jovenes se acompaña en la visita.

"These young persons will be accompanied during the visit."

The first step towards simplification is the identification of a target structure. Here we look for a syntax sub-tree where a relative pronoun is pre-verbal and directly or indirectly depends on an inflected verb. A target structure for this simplification type is shown in Fig. 3. In the next step, the relative pronoun is replaced by the full noun form of its antecedent. In some cases a matching article has to be added. Finally, the dependency branch between the main verb of the relative clause and the superordinate clause is cut, resulting in two independent clauses, as shown in Fig. 4.

The simplification operations shown here are implemented as syntactic rules within the MATE framework (Bohnet et al. 2000), mentioned in Sect. 5, and they operate on the output of a dependency parser (Bohnet 2009). The syntactic rules in

[20] Light-to-right order in the tree does not necessarily represent linear order of words in the sentence.

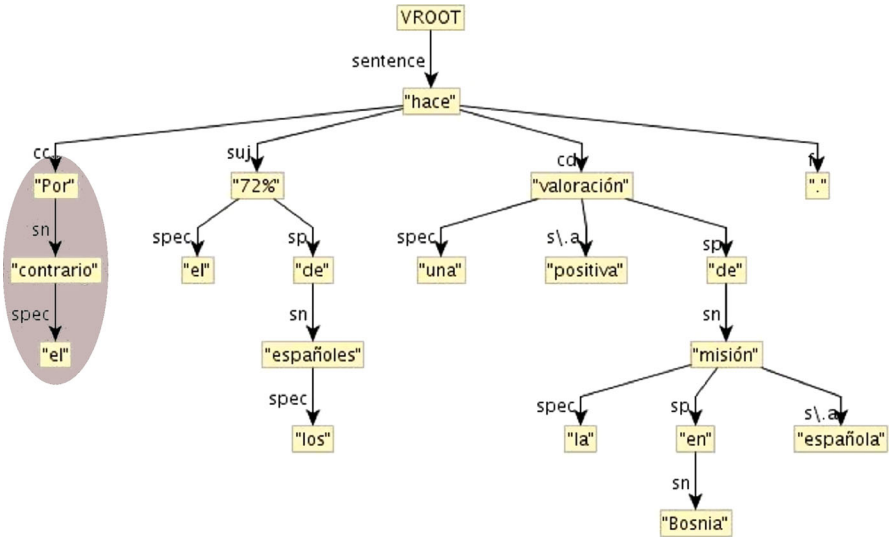

Fig. 1 A syntactic dependency tree before simplification

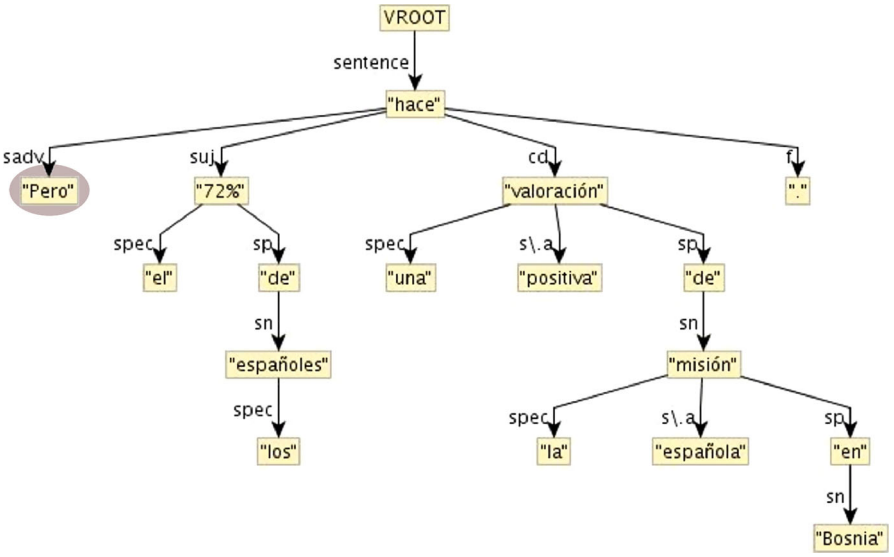

Fig. 2 A syntactic dependency tree after simplification

MATE identify a target structure in what is called the *left side* of the rule and map them to new nodes and relations on the *right side* of the same rule. In addition, conditions can be introduced on the left side, for example a condition that a node has to correspond to a certain part of speech or that it has to carry certain inflectional

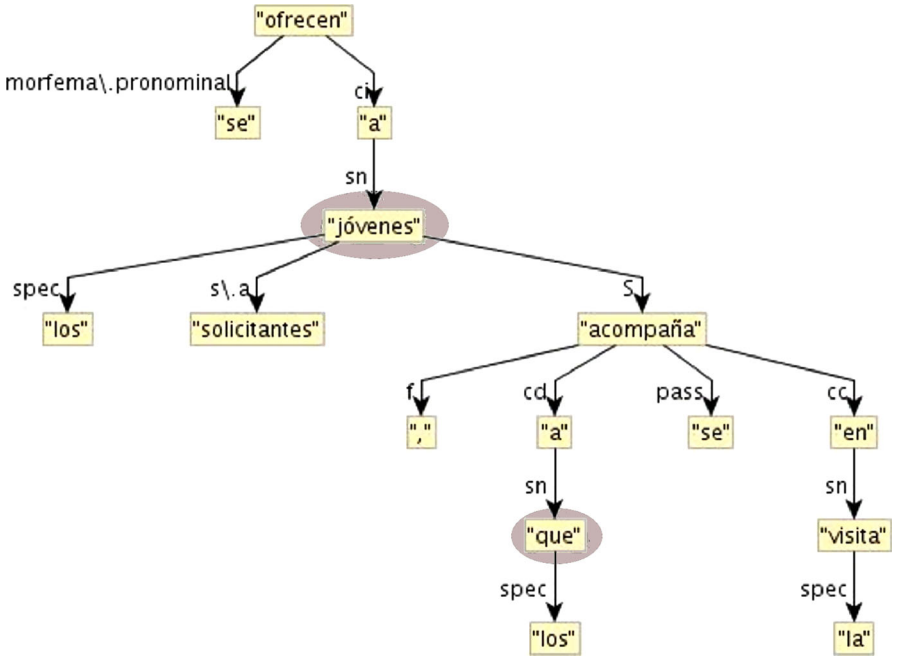

Fig. 3 A target structure containing a relative clause

information. Table 3 shows the rule which operates on Fig. 1. It looks for a tree fragment (specified in the left side) with three nodes, for which the conditions in the condition set must hold, namely that they have the lexical content of *por, lo* and *contrario*, respectively. Then it maps these nodes to a newly created tree fragment on the right side of the rule. The <=> operator establishes equivalence between nodes on the left side and the right side of the rule. Note that this rule will not produce the tree in Fig. 2 directly, because it creates nodes without lexical content (marked as delete = yes), which have to be deleted by a further clean-up rule.

The manipulation of (14) requires a larger set of rules which, expressed informally, require three nodes: a noun, a verb and a relative pronoun. A sentential (subordinating) relation must hold between the noun and the verb. Then three corresponding nodes in the output structure are created, corresponding to the three input nodes. The space between the noun and the subordinate verb is then marked as the cut-off point. Finally, the relation which still holds between the matrix clause and the former relative clause is cut in the subsequent step.

Note that there is also a *change* operation involved in the treatment of this example. The relative pronoun has to be substituted by a lexical NP. In this case the lexical content of the noun can be copied in place of the former relative pronoun. This only involves the copying of lexical content from one syntactic node into another, but in some cases the presence of an article must be checked, or a tree fragment must be copied from the matrix clause into the place of the relative pronoun. Note that, even if not strictly necessary in the case of relative pronouns,

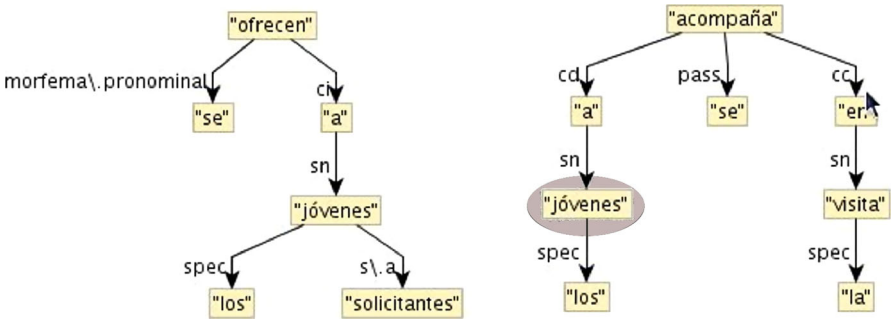

Fig. 4 The relative clause converted into an independent sentence

Table 3 The syntactic rule which converts *por lo contrario* in Fig. 1 to *pero*

Left side	Right side	Conditions

?Xl {
sn− > ?Yl {
spec− > ?Zl
}
}
	rc:?Xr {
	<=> ?Xl
	slex=""
	rc:?r− > rc:?Yr {
	<=> ?Yl
	slex="Pero"
	ppos=cc
	rc:?r2− > rc:?Z2r{
	<=> ?Zl
	slex=""
	delete=yes
	}
	}
	}
		?Xl.slex="por"
		or ?Xl.slex="Por";
		?Yl.slex="contrario,"
		or ?Yl.slex="contrario";
		?Zl.slex="el";

many change operations which copy material from an antecedent NP need information on co-reference.

It can be seen from our examples that the approach we take is largely based on a hand-crafted grammar. This is a necessary consequence of the fact that there are no large parallel corpora of simplified text available for Spanish. The grammar is developed in a work cycle which typically involves three steps: First a rule is written on the basis of selected examples, and then the grammar is applied to a development corpus. All applications of a rule are manually inspected and rules are refined to either avoid erroneous rule applications or to correct the output.

We plan to complement this approach with data-driven approaches to sub-problems of text simplification, for example in the decision if a certain construction needs to be simplified or not. Also the implicit word sense disambiguation which is necessary in the case of single word substitution appears to require a statistical support system for the selection of synonyms (Bott et al. 2012).

An evaluation of the performance of the different simplification operations is given in Table 4. This evaluation was carried out over 886 sentences, taken from the

Table 4 Percentage of right rule application and frequency of application (percentage of sentences affected) per rule type

Operation	Precision	Recall	Frequency (%)
Relative clauses (all types)	0.393	0.661	20.65
Simple relative clauses	0.371	–	19.18
Complex relative clauses	0.692	–	0.90
Participle and gerundive constructions	0.636	0.206	2.48
Object coordination	0.420	0.583	7.79
VP and clause coordination	0.648	0.500	6.09

part of the corpus which is only composed of original unsimplified texts and which was not used as part of the development set. For the calculation of recall we manually annotated 262 sentences for structures which contain a target structure that could be simplified. We applied our simplification grammar to these texts and annotated the output, counting the places where the rule had produced a felicitous output while ignoring minor grammaticality issues which could be solved with further fine-tuning of the grammar rules.[21].

(i) a. Wikileaks es una página web. [En] este web se da información sobre asuntos de interés público.

Wikileaks is a web page. [On] this web page information of public interest is given.

b. … estas ONG instan a la OTAN a tomar medidas urgentes … [Este] OTAN celebra … una cumbre en Lisboa.

… these ONGs ask the NATO to take urgent measures … [This] NATO celebrates a summit in Lisbon.

The precision here is defined as the ratio between all applications and correct applications of each rule. The frequency of rule application is given as the percentage of sentences affected by a rule. We distinguished between split operations which operate on two types of relative clause constructions[22] (depending on whether they are headed by a preposition or not), participle and gerundive constructions (like example (8)), and two types of coordination constructions.

In interpreting Table 4 it is important to note that parse errors are a serious problem and propagate into the simplification module. These constitute a large part of all errors, up to 37 % in the category of participle constructions. Error analysis showed us that there is still much room for improvement of precision and recall with further grammar engineering.

[21] Such errors included cases where a preposition required by the construction was missing (such as *a/in* in the case of (i-a)) or an article was wrongly inserted before a proper name (as *este/this* in (i-b)).

The errors which were encountered were later corrected by improving the rules. The sentences used for rule improvement are excluded from the test set for future evaluations.

[22] The annotation scheme did not allow us to calculate recall for the two different relative clause splitting operations separately and these values are not listed in Table 4.

In the current version of the simplification prototype, we have concentrated on two simplification operations which we considered crucial and representative: lexical substitution and sentence splitting. Lexical substitution is the single most frequent simplification operation we could observe and it has an important influence on the difficulty or ease with which texts can be understood. This operation can be handled gracefully as an operation on syntactic trees, even if the unit to be substituted is not larger than one single word. More importantly, this approach can be extended to discontinuous lexical units and to cases where a lexical unit is substituted by a different lexical unit which also occupies another position in the sentence (as, for example, in the case where intra-sentential *sin embargo* (*however*) is substituted by a sentence initial *pero* (*but*)). Sentence splitting is also a relatively frequent operation and it has a very important impact on text complexity, namely on the number of words per sentence and the embedding depth of the syntactic tree. Sentence splitting operations also resemble other syntactic change, and insertion operations and similar technique can be applied in those cases.

8 Conclusions and outlook

In this paper we presented the first steps towards the creation of an automatic text simplification system for Spanish. Text simplification has a wide range of applications and target users, from language learners and foreign residents in a new country to elderly people and people with cognitive difficulties. At present, there is, to the best of our knowledge, no automatic text simplification system for Spanish. Since manual simplification involves a large amount of manual work, an automatic tool for this language has a large list of potential uses.

We presented a corpus-based study that surveys the different operations that an automatic text simplification system must be able to carry out. We have identified a series of operations that human editors typically carry out and the frequency with which these occur. Some of these operations involve cognitive capacities which only humans possess and which are beyond the possibilities of a computational system. On the other hand, we could find a series of simplifying operations which are well defined and can be properly modeled as rules that operate on syntactic trees. On the basis of our findings, we started to develop a simplification prototype, which is still in a development phase. The most important purpose of this prototype in its present form is to demonstrate that some operations can be handled successfully, especially operations which are frequent and which have an important influence on text readability.

In future work we will refine the syntactic simplification module by developing a broad-coverage grammar. A global evaluation with human participants from the target group is in preparation and we will report on it in the near future. We will also amplify the corpus study presented in Sect. 6 when we have the full corpus or, at least a larger part of the corpus available, in order to study if the frequency of edit operations are stable across text domains. Since in the case of syntactic simplification we could not find a large enough data set to test data-driven techniques, we want to explore the possibility to use a hybrid approach, in which

some of the most frequent simplification rules can be learned automatically, which can then be manually polished and complemented by hand-crafted rules.

We are also interested in the development of reliable metrics for text complexity, both lexical and structural. Such metrics are important for the development process, since they can indicate the progress at intermediate points of the system development, and manual evaluation is too costly to carry out repeatedly.

Acknowledgments We are grateful to five anonymous reviewers for their very constructive comments and insights which helped us improve the final version of the paper. We would also like to thank Simon Mille for his substantial help with the MATE grammar framework. The research described in this paper arises from a Spanish research project called 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 of Scientific Research, Development and Technological Innovation (I+D+i), within the strategic Action of Telecommunications and Information Society (Avanza Competitiveness, with file number TSI-020302-2010-84). We are grateful to the fellowship RYC-2009-04291 from Programa Ramón y Cajal 2009 and to the project SKATER-UPF-TALN (TIN2012-38584-C06-03), Ministerio de Economía y Competitividad, Secretaría de Estado de Investigación, Desarrollo e Innovación, Spain. We are grateful to Biljana Drndarevic for proofreading the paper.

References

Aluísio, S. M., Specia, L., Pardo, T. A. S., Maziero, E. G., & de Mattos Fortes, R. P. (2008). Towards Brazilian Portuguese automatic text simplification systems. In *ACM symposium on document engineering* (pp. 240–248).

Anula, A. (2007). Tipos de textos, complejidad lingüística y facilicitación lectora. In Man-Ki, Jy-Eun, y Macas (Eds.), *Actas del Sexto Congreso de Hispanistas de Asia* (pp. 45–61). República de Corea: Seúl.

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

Anula, A. (2011). *Pautas básicas de simplificación textual y diseño del corpus SIMPLEXT. Technical report, Grupo DILES*. Madrid, Spain: Universidad Autónoma de Madrid.

Aranzabe, M., de Ilarraza, A., & Gonzalez-Dios, I. (2012). First approach to automatic text simplification in Basque. In *Natural language processing for improving textual accessibility (NLP4ITA) workshop programme* (pp. 1–8).

Barthe, K., Juaneda, C., Leseigneur, D., Loquet, J.-C., Morin, C., Escande, J., et al. (1999). GIFAS rationalized French: A controlled language for aerospace documentation in French. *Technical Communication, 46*(2), 220–229.

Barzilay, R., & Elhadad, N. (2003). Sentence alignment for monolingual comparable corpora. In *Proceedings of the 2003 conference on empirical methods in natural language processing* (pp. 25–32).

Bohnet, B. (2009). Efficient parsing of syntactic and semantic dependency structures. In *Proceedings of the conference on natural language learning (CoNLL)* (pp. 67–72). Boulder, Colorado: Association for Computational Linguistics.

Bohnet, B., Langjahr, A., & Wanner, L. (2000). A development environment for an MTT-based sentence generator. In *Proceedings of the first international conference on natural language generation* (pp. 260–263). Stroudsburg, PA, USA: Association for Computational Linguistics.

Bott, S., Rello, L., Drndarević, B., & Saggion, H. (2012). Can Spanish be simpler? LexSiS: Lexical Simplification for Spanish. In *Proceedings of Coling 2012: The 24th International Conference on Computational Linguistics*.

Bott, S., & Saggion, H. (2011). An unsupervised alignment algorithm for text simplification corpus construction. In *Workshop on monolingual text-to-text generation, co-located with ACL 2011* Porland, Oregon.

Bouayad-Agha, N., Casamayor, G., Ferraro, G., & Wanner, L. (2009). Simplification of patent claim sentences for their paraphrasing and summarization. In *FLAIRS Conference*.

Brown, K. (1995). Current Issues in Plain English. *ARIS Bulletin, 6*(4).

Canning, Y., Tait, J., Archibald, J., & Crawley, R. (2000). Cohesive generation of syntactically simplified newspaper text. In *Proceedings of the third international workshop on text, speech and dialogue* (pp. 145–150).

Carroll, J., Minnen, G., Canning, Y., Devlin, S., & Tait, J. (1998). Practical simplification of English Newspaper text to assist aphasic readers. In *Proceedings of AAAI-98 workshop on integrating artificial intelligence and assistive technology* (pp. 7–10).

Chandrasekar, R., Doran, C., & Srinivas, B. (1996). Motivations and methods for text simplification. In *Proceedings of the international conference on computational Linguistics* (pp. 1041–1044).

Coster, W., & Kauchak, D. (2011). Learning to simplify sentences using Wikipedia. In *Proceedings of the workshop on monolingual text-to-text generation* (pp. 1–9). Portland, Oregon: Association for Computational Linguistics.

Crossley, S. A., & Mcnamara, D. S. (2008). Assessing L2 reading texts at the intermediate level: An approximate replication of Crossley, Louwerse, McCarthy & McNamara (2007). *Language Teaching, 41*(03), 409–429.

Daelemans, W., Hthker, A., & Sang, E. T. K. (2004). Automatic sentence simplification for subtitling in Dutch and English. In *Proceedings of the 4th conference on language resources and evaluation* (pp. 1045–1048). ELRA.

De Belder, J., Deschacht, K., & Moens, M. (2010). Lexical simplification. In *Proceedings of ITEC2010: 1st international conference on interdisciplinary research on technology, education and communication*.

Dell'Orletta, F., Montemagni, S., & Venturi, G. (2011). Read-it: Assessing readability of italian texts with a view to text simplification. In *Proceedings of the second workshop on speech and language processing for assistive technologies* (pp. 73–83). Association for Computational Linguistics.

Devlin, S., & Tait, J. (1998). The use of a psycholinguistic database in the simplification of text for aphasic readers. *Linguistic Databases*, 161–173.

DuBay, W. (2004). The principles of readability. *Impact Information*, 1–76.

Feng, L., Elhadad, N., & Huenerfauth, M. (2009). Cognitively motivated features for readability assessment. In *EACL* (pp. 229–237).

Feng, L., Jansche, M., Huenerfauth, M., & Elhadad, N. (2010). A comparison of features for automatic readability assessment. In *Proceedings of the international conference on computational Linguistics (Posters)* (pp. 276–284).

Flesch, R. (1948). A new readability yardstick. *Journal of applied psychology, 32*(3), 221–233.

Gasperin, C., Maziero, E. G., & Aluísio, S. M. (2010). Challenging choices for text simplification. In *PROPOR* (pp. 40–50).

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods Instruments Computers a Journal of the Psychonomic Society Inc, 36*(2), 193–202.

Hyönä, J., & Olson, R. (1995). Eye fixation patterns among dyslexic and normal readers: Effects of word length and word frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(6), 1430.

Inui, K., Fujita, A., Takahashi, T., Iida, R., & Iwakura, T. (2003). Text simplification for reading assistance: A project note. In *Proceedings of the second international workshop on Paraphrasing—volume 16, PARAPHRASE '03* (pp. 9–16). Stroudsburg, PA, USA: Association for Computational Linguistics.

Jing, H. (2002). Using Hidden Markov Modeling to decompose human-written summaries. *Computational Linguistics, 28*, 527–543.

Klebanov, B. B., Knight, K., & Marcu, D. (2004). Text simplification for information-seeking applications. In *On the move to meaningful internet systems, lecture notes in computer science* (pp. 735–747). Berlin: Springer.

Krifka, M. (2007). Approximate interpretation of number words: A case for strategic communication. *Cognitive Foundations of Interpretation*, 111–126.

Max, A. (2006). Writing for language-impaired readers. In *Proceedings of the conference on intelligent text processing and computational Linguistics* (pp. 567–570).

Maynard, D., Tablan, V., Cunningham, H., Ursu, C., Saggion, H., Bontcheva, K., et al. (2002). Architectural elements of language engineering robustness. *Journal of Natural Language*

Engineering – Special Issue on Robust Methods in Analysis of Natural Language Data, 8(2/3), 257–274.

Mille, S., & Wanner, L. (2008). Making text resources accessible to the reader: The case of patent claims. In *Proceedings of the language resources and evaluation conference*, Marrakech (Marocco).

Ogden, C. K. (1937). *Basic English: A general introduction with rules and grammar*. London: Paul Treber.

Ong, E., Damay, J., Lojico, G., Lu, K., & Tarantan, D. (2008). Simplifying text in medical literature. *Journal of Research in Science, Computing and Engineering, 4*(1).

Padró, L., Collado, M., Reese, S., Lloberes, M., & Castellón, I. (2010). FreeLing 2.1: Five years of open-source language processing tools. In *Proceedings of 7th language resources and evaluation conference* Malta: La Valletta.

Petersen, S. E., & Ostendorf, M. (2007). Text simplification for language learners: A corpus analysis. In *Proceedings of workshop on speech and language technology for education*.

Petz, A., & Tronbacke, B. (2008). People with specific learning difficulties: Easy to read and HCI. In *ICCHP* (pp. 690–692).

Pitler, E., & Nenkova, A. (2008). Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the conference on empirical methods in natural language processing, EMNLP '08* (pp. 186–195). Stroudsburg, PA, USA: Association for Computational Linguistics.

Power, R., & Williams, S. (2011). Generating numerical approximations. *Computational Linguistics, 38*(1), 113–134.

Rodríguez Diéguez, J., Moro Berihuete, P., & Cabero Pérez, M. (1992). La predicción de la lecturabilidad de los textos escritos. In *X Congreso Nacional de Pedagogía* Salamanca.

Saggion, H. (2008). Automatic summarization: An overview. *Revue française de linguistique appliquée, XIII*(1).

Saggion, H., Gmez-Martnez, E., Etayo, E., Anula, A., & Bourg, L. (2011). Text simplification in simplext: Making text more accessible. *Revista de la Sociedad Espaola para el Procesamiento del Lenguaje Natural, 47*, 341–342.

Seretan, V. (2012). Acquisition of syntactic simplification rules for French. In Chair), N. C. C., Choukri, K., Declerck, T., Dogan, M. U., Maegaard, B., Mariani, J., Odijk, J., & Piperidis, S. (Eds.), *Proceedings of the eight international conference on language resources and evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA).

Siddharthan, A. (2002). An architecture for a text simplification system. In *Proceedings of the language engineering conference (LEC'02)* (pp. 64–71).

Siddharthan, A. (2011). Text simplification using typed dependencies: A comparison of the robustness of different generation strategies. In *Proceedings of the 13th European workshop on natural language generation (ENLG)* (pp. 2–11).

Specia, L. (2010). Translating from complex to simplified sentences. In *PROPOR* (pp. 30–39).

Specia, L., Jauhar, S. K., & Mihalcea, R. (2012). SemEval-2012 task 1: English lexical simplification. In *Proceedings of the first joint conference on lexical and computational semantics—volume 1: Proceedings of the main conference and the shared task, and volume 2: Proceedings of the sixth international workshop on semantic evaluation, SemEval '12* (pp. 347–355). Stroudsburg, PA, USA: Association for Computational Linguistics.

Tanguy, L., & Tulechki, N. (2009). Sentence complexity in French: A corpus-based approach. *Intelligent information systems (IIS)*, pages 1–14.

Watanabe, W. M., Junior, A. C., de Uzêda, V. R., de Mattos Fortes, R. P., Pardo, T. A. S., & Aluísio, S. M. (2009). Facilita: Reading assistance for low-literacy readers. In *SIGDOC* (pp. 29–36).

Woodsend, K., & Lapata, M. (2011). Learning to simplify sentences with quasi-synchronous grammar and integer programming. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 409–420). Association for Computational Linguistics.

Zhu, Z., Bernhard, D., & Gurevych, I. (2010). A monolingual tree-based translation model for sentence simplification. In *Proceedings of The 23rd international conference on computational linguistics* (pp. 1353–1361). Beijing, China.