Empirical Identification of Text Simplification Strategies for Reading-Impaired People

Susana BAUTISTA [a] Carlos LEÓN [a] Raquel HERVÁS [a] and Pablo GERVÁS [a]
[a] *Departamento de Ingeniería del Software e Inteligencia Artificial, Universidad Complutense de Madrid, Spain*
subautis@fdi.ucm.es, cleon@fdi.ucm.es, raquelhb@fdi.ucm.es, pgervas@sip.ucm.es

Abstract.

Objective The way in which information is written or presented can exclude many people, especially those who have problems to read, write or understand. However, the process of simplifying texts by hand is extremely time and effort consuming so any attempts to automate part of this process can leverage the access to information. The final objective of this research is the development of an automatic system that eases the process of text adaptation for reading-impaired people.

Main Content In this paper we present a preliminary analysis of a parallel corpus containing simplified texts along with their original versions in order to identify what kind of transformations are used when creating the easy-to-read version of a text. The study has been carried out to find out how this process can be automated, at least partially.

Results The study has evidenced that a parallel corpus containing easy-to-read versions of texts becomes a very powerful source of knowledge. First, it provides very important information regarding the type of operations carried out during simplifications, and the relative frequencies of each type of operation. Second, it constitutes a valuable resource for identifying empirical instances of the various types of transformations, which can be used as data during development or even training for an automatic or semi-automatic system to simplify texts. Both the frequencies and the types of transformations have been extracted from the study of the corpus. We therefore propose this analysis as a basic resource for the development of an automatic simplification system.

Conclusion In this paper we have presented a preliminary analysis of the type of operations involved in text simplification. This first step is previous to any attempt to automatize of the process, as it is required to know what kind of phenomena can be expected. We can conclude that the first step towards a useful formalization of how to simplify texts has been taken.

Keywords. Text simplification, parallel corpora, paraphrases, accessibility

Introduction

Literacy is the combination of reading, writing, speaking and listening skills that we all need to fulfill. These skills are essential to happiness, health and wealth of individuals and society. The most recent surveys of literacy in the UK [1,2,3] reveal that 7 million adults in England –one in five adults– cannot locate the reference page for plumbers if given the Yellow Pages alphabetical index. That is an example of functional illiteracy. It means that one in five adults has less literacy than the expected literacy in an 11-year-old child.

According to Snow [4] reading comprehension entails three elements: the *reader* who is meant to comprehend; the *text* that is to be comprehended, and the *activity* in which comprehension takes part. In addition to the content presented in the text, the vocabulary load of the text and its linguistic structure, discourse style, and genre interact with the reader's knowledge and have an influence on reading comprehension too. When these factors do not match the reader's knowledge and experience, the text becomes too complex for an appropriate comprehension.

Text simplification aims at providing human readers with a better understanding of a written text through its simplification. It is known that long, conjoined or passive sentences, embedded clauses, non-canonical word order, and use of low-frequency words, among other things, increase text complexity for language-impaired readers [5,6,7,8,9]. There are different initiatives that propose guidelines that may help when rewriting a text to make it more comprehensive. Some of them are *Plain Language* [10], the *European Guidelines for the Production of Easy-to-Read Information* [11] and the *Web Content Accessibility Guidelines* [12].

However, little work has been devoted to how these general guidelines can be directly applied in the text simplification process. In this paper we present the preliminary analysis of a parallel corpus generated by experts in which original and simplified sentences are aligned. While experts are able to generate simplified sentences, they seldom explicitly identify the set of rules they are applying. This paper reports the identification of transformations used by experts to obtain easy-to-read versions of texts in the corpus. The ultimate aim is to develop a set of rules to suggest automatically possible transformations to assist future transcriptions.

1. Previous Work

Text simplification systems can be studied along three axes: the type of system -rule-based or corpus-based-, the type of knowledge used to identify the need for simplification, and the goals of the system.

A few rule-based systems have been developed for text simplification [13,14,8], focusing on different kinds of readers (poor literate, aphasic, etc). These systems contain a set of manually created simplification rules that are applied to each sentence, being these rules usually based on parsed structures and limited to certain simplification operations. [14] proposes a syntactic simplification architecture that relies on shallow text analysis and favors time performance. The general goal of the architecture is to make texts more accessible to a broader audience. [15] applies text simplification in the writing process by embedding an interactive text simplification system into a word processor. At

the user's request, an automatic parser analyzes an individual sentence and the system applies handcrafted rewriting rules. This system requires human intervention in every step. [16] propose a rule-based system for text simplification aimed at deaf people. In Easy2Read Standard [1]some rules for eAccessibility in websites are recommended.

The transformation of texts into easy-to-read versions can be phrased as a translation problem between two different subsets of language: the original and the easy-to-read version. Corpus-based systems can learn from corpora the relevant simplification operations and also the required degree of simplification for a given task. Pettersen and Ostendorf [17] address the task of text simplification in the context of second-language learning. A data-driven approach to simplification is proposed using a corpus of paired articles in which each original sentence does not necessarily have a corresponding simplified sentence, making it possible to learn where writers have dropped or simplified sentences. A classifier is used to select the sentences to simplify, and Siddharthan's syntactic simplification system is used to split the selected sentences.

Some language technology systems attempt to simplify documents for various purposes. A variety of simplification techniques, like substituting common words for uncommon words [18], activating passive sentences and resolving references [19], reducing multiple-clause sentences to single-clause sentences [20,19,5] and making appropriate choices at the discourse level [21] have been used.

2. Analysis of Simplifications in the Corpus

Our aim is to study what kind of simplification transformations are involved in a parallel corpus of easy-to-read versions of texts aligned with the original sources.

2.1. Corpus

The study is based on a parallel corpus created by Barzilay and Elhadad [22]. The corpus is a compilation of two aligned collections from the Encyclopedia Britannica and Britannica Elementary. In contrast to the long detailed articles of the Encyclopedia Britannica, Britannica Elementary contains one-to-two page entries targeted to children. The elementary version generally contains a subset of the information presented in the original version, but there are cases in which the elementary entry contains additional or more up-to-date pieces of information. There are over 2,600 easy-to-read articles designed to help elementary students (for ages 6-10) thrive in school.

The Britannica Elementary has not been created for people with literacy problems. While authors acknowledge that there exist important differences between different types of transformation, this approach is considered an acceptable preliminary study.

Part of the original corpus was aligned by hand to provide a core subset for training and testing. The present study was centered on the fragment of this human-alligned subset used for training, 11 texts with a total of 320 sentences, which comprise the 37.93% of the whole corpus. In each text the sentences of the original source (the Encyclopedia Britannica version) are alligned with the corresponding sentence from the simplified version (the Britannica Elementary version).

Table 1 shows an example of a fragment of one of these texts, organized so the pieces of text that correspond to the same meaning are presented in the same row.

[1] http://www.informationarchitects.jp/en/100e2r/ [11/5/2011]

Table 1. Example of transformations in an aligned text from the corpus

Original text	Simplified text
Ancient Egyptians believed that	
Osiris, a good and wise king,	King Osiris
was the first pharaoh.	was the first pharaoh
He spread knowledge	and spread knowledge.
to other parts of the world,	
while	
his wife, Isis,	Queen Isis
ruled Egypt	ruled Egypt
in his place.	when he was gone.

2.2. Sentence Level Study

The example shown in Table 1 shows that the correspondence between sentences across the two versions is not 1-to-1, but n-to-m (a set of sentences can be simplified in only one sentence, or vice versa). To cover this problem the sentences have been aligned by hand in the first stage of the processing. For each pair of aligned sentences, a second level of alignment is carried out between their lexical and syntactical structure.

Once the structures of sentences in the two versions have been aligned, the second stage of the study involves the explicit identification of the transformations that lead from the original to the simplified version. The following types of transformations have been identified:

- Lexical transformations (the use of synonyms, the replacement of words with easy-to-read alternatives).
- Syntactic transformations that do not affect the semantics. For instance, sentences with two concepts that are aligned with two simpler sentences.
- Deletion of non-relevant information.
- Addition of extra information in the simplified version used to better explain difficult concepts.
- Complete rewrite of the original sentence (paraphrase).

2.3. Taxonomy

Once the whole study was finished and the number of occurrences of each type was measured, a taxonomy of transformations was created and the frequency of each one was analysed. At the top of the taxonomy a distinction is made between transformations that modify the CONTENT of the sentence, and those that only modify the FORM or structure of the sentences but respect their meaning. This latter group is referred to as paraphrases. Two kinds of paraphrases are considered: LEXICAL paraphrases where modifications are restricted to word substitution, and SYNTACTIC paraphrases where modifications involved altering the syntax of the sentence in some way. From of the point of view of CONTENT, we have considered two kinds: ADDITION or DELETION. This classification is consistent with current views expressed in the literature [23,24], though there is no currently agreed canonical taxonomy of paraphrases. Some examples of syntactic paraphrase are presented in Table 2.

Table 2. Examples of paraphase

Type	Original text	Simplified text
Lexical	It is believed that 30,000 lives were lost	It is believed that 30,000 people died
Lexical	It is the most fashionable district of Budapest, where Hungary's elite have houses	It is the most fashionable district of Budapest, where Hungary's wealthy have houses
Syntactic	Madrid was occupied by French troops during the early 19th century	France occupied Madrid during the early 19th century
Syntactic	Baghdad has become an active cultural center for the Arab world	Baghdad is a center of Islamic art

Table 3. Results of preliminary analysis

Transformation		Num. of Transf.
CONTENT	Deletion	230 (51.34%)
	Addition	48 (10.71%)
FORM	Paraphrasing	170 (37.95%)
TOTAL TRANSF.		448

Table 4. Results of the analysis of paraphrases

Type of paraphrases		Number of Transformations	Percentage
LEXICAL	Noun synonym	36	21.17%
	Adjective synonym	19	11.18%
	Verb synonym	35	20.59%
TOTAL LEXICAL		90	52.94%
SYNTACTIC	Passive to active	11	6.47%
	Perfect to simple tense	19	11.18%
	Sentence structure	50	29.41%
TOTAL SYNTACTIC		80	47.06%
TOTAL PARAPHRASES		170	100%

2.4. Results

Table 3 shows the results of our analysis. The majority of transformations (51.34 %) are deletions of information to generate the simplified version. Paraphrases come next (37.95 %), followed by instances of addition of clarifying information (10.71 %). Notice also that there are 448 transformations in 320 sentences, because usually more than one transformation is applied in each sentence.

The results of a close analysis of paraphrase are presented in Table 4. It seems that lexical and syntactic paraphrases appear in the corpus with a similar proportion, with a slight preference for lexical ones (52.94 % to 47.06 %). Transformations of sentence structure are specially frequent (29.41 %).

3. A Proposal for an Automated Simplification System

A parallel corpus containing easy-to-read versions of texts becomes a very powerful source of knowledge. First, it provides very important information regarding the type of operations carried out during simplifications, and the relative frequencies of each type of operation. Second, it constitutes a valuable resource for identifying empirical instances of the various types of transformations, which can be used as data during development or even training for an automatic or semi-automatic system to simplify texts.

The general schema for the subsequent process is planned as follows:

1. A subset of the operations identified above is selected, based on their relative frequency of use and the availability of computational resources for implementing them.
2. For the specific set of chosen operations, an empirical process of rule extrapolation is carried out over the set of correspondences for the relevant operation contained in the corpus.
3. A computational implementation of these rules is developed.
4. The resulting module is applied to the source texts corresponding to the part of the corpus set aside for testing, and results are compared to the associated output versions.

Each of the identified types of transformations would have to be treated differently.

The most used transformation by humans is the deletion of information from the original sentence to generate the easy-to-read version. Deletion is complex because deciding what to delete involves a very complex analysis of the semantics. And this can not be done always. One of the main aspects to consider in this operation is the context and the meaning of the sentence. Human knowledge is way more powerful than machines to determine which parts can be safely deleted and which parts can not. A first possible step will could be to apply a summarization methodology to discard not essential information. Later we could apply our simplification methodology to adapt the text.

For lexical paraphrases, we are considering the use of WordNet [25] as a source of possible word substitutions. Preliminary work on this approach [8] has shown that WordNet provides a broad range of options for substitution. It is still necessary to identify which are the most useful criteria to apply: whether the words that should be chosen are more frequently used synonyms for the original word, less ambiguous synonyms for it, more specific hyponyms when the original word is too abstract, or more abstract hypernyms when the original word is too specific.

For the automation of syntactic paraphrase we are considering the implementation of a set of rules for the transformation of parse trees obtained automatically. As tools for the automated analysis of sentences we are considering the use of the Stanford Parser [26] (for constituent-based analysis) and MINIPAR [27] (for dependency-based analysis).

For transformations involving the additional of information we are considering the use of WordNet glosses of those words corresponding to less frequent uses of WordNet concepts. Again, the context and meaning of the original sentence play a fundamental role, and this kind of information has to be included in the adaptation rules that will be defined and applied in the automatic system. This study can help to decide whether addition must be performed or not.

4. Conclusions and Future Work

Using an annotated and aligned corpus previously available, in this paper we have presented a preliminary analysis of the type of operations involved in text simplification. This first step is previous to any attempt of automating the process, as it is required to know what kind of phenomena can be expected.

The final aim of our work is to gather more knowledge in order to build a system for the semi-automatic simplification of texts by suggesting text transformations that are afterwards approved by an expert. Among the different types of transformations, deletions and paraphrases seem very promising because there are different systems that deal with the recognition and modification of sentences at syntactic and lexical levels.

Our work is developed with an aligned and annotated corpus in English. These ideas could be applied to any language which had a corpus aligned and annotated in a similar way.

Preserving the meaning of the original sentence and generating a easy-to-read version following the guidelines is a challenge when rewriting the whole sentence. This part is not addressed in this paper, but it is one of the most interesting application of this study, and it is included as part of the future work.

We are planning to explore the use of other similar corpora. A good candidate is the use of the Inclusion Europe web pages [28]. Inclusion Europe is the European Association of Societies of Persons with Intellectual Disabilities and their Families. In their website they publish two versions of each page: the original one and an easy-to-read version. From these pages we can build a parallel corpus that can be studied and used as the Encyclopedia Britannica one. However, a first step of alignment between the two versions must be performed before any other kind of analysis is possible.

The potentialities of text simplification systems for education, for example, are evident. For students, it is a first step for more effective learning. For people with poor literacy, we see text simplification as a first step towards social inclusion, facilitating and developing reading and writing skills to interact in society. The social impact of text simplification is undeniable. This makes the automation of text simplification an interesting field of research.

References

[1] G. D. Deeqa Jama, "Literacy: State of the nation," National Literacy Trust, Tech. Rep., 2010.

[2] D. for Education and S. from United Kingdom, "Skills for life," Tech. Rep., 2003.

[3] D. J. Clark Christina, "Young people reading and writing today: Whether, what and why," London: National Literacy Trust, Tech. Rep., 2010.

[4] C. E. Snow, U. States., Science, and T. P. I. R. Corporation), *Reading for understanding : toward an R&D program in reading comprehension / Catherine Snow*. Rand, Santa Monica, CA :, 2002.

[5] A. Siddharthan, "Resolving attachment and clause boundary amgiguities for simplifying relative clause constructs," in *Proceedings of the Student Research Workshop, 40th Meeting of the Association for Computacional Linguistics*, 2002.

[6] B. B. Klebanov, K. Knight, and D. Marcu, "Text simplification for information-seeking applications," in *On the Move to Meaningful Internet Systems, Lecture Notes in Computer Science*. Springer Verlag, 2004, pp. 735–747.

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

[8] S. Bautista, P. Gervás, and R. Madrid, "Feasibility analysis for semiautomatic conversion of text to improve readability," in *Proceedings of the Second International Conference on Information and Communication Technology and Accessibility*, May 2009 2009.

[9] H. M. Caseli, T. F. Pereira, L. Specia, T. A. S. Pardo, C. Gasperin, and S. M.Aluisio, "Building a Brazilian Portuguese Parallel Corpus of Original and Simplified Texts," in *In Proceedings of CICLing*, 2009.

[10] The Plain Language Action and Information Network (PLAIN), "Plain Language," *http://www.plainlanguage.gov* [17/9/2010], 2005.

[11] G. Freyhoff, G. Hess, L. Kerr, E. Menzel, B. Tronbacke, and K. V. D. Veken, "European guidelines for the production of easy-to-read information," *http://www.osmhi.org/contentpics/139/ European Guidelines for ETR publications (2).pdf* [17/9/2010], 1998.

[12] W3C, "Web content accessibility guidelines," *http://www.w3.org/TR/WCAG20/* [7/2/2011], 2008.

[13] R. Chandrasekar, C. Doran, and B. Srinivas, "Motivations and methods for text simplification," in *In Proceedings of the Sixteenth International Conference on Computational Linguistics (COLING '96)*, 1996, pp. 1041–1044.

[14] A. Siddharthan, "Syntactic Simplification and Text Cohesion," Ph.D. dissertation, Research on Language and Computation, 2003.

[15] A. Max, "Writing for language-impaired readers," in *CICLing*, 2006, pp. 567–570.

[16] K. Inui, A. Fujita, T. Takahashi, R. Iida, and T. Iwakura, "Text simplification for reading assistance: a project note," in *Proceedings of the second international workshop on Paraphrasing*. Morristown, NJ, USA: Association for Computational Linguistics, 2003, pp. 9–16.

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

[18] S. Devlin and J. Tait, *Linguist Databases*. CSLI, 1998, ch. The use of a Psycholinguistic database in the Simplification of Text for Aphasic Readers, pp. 161–173.

[19] Y. Canning, "Cohesive simplification of newspaper text for aphasic readers," in *3rd annual CLUK Doctoral Research Colloquium*, 2000.

[20] R. Chandrasekar and B. Srinivas, "Automatic induction of rules for text simplification," *Knowledge-Based Systems*, vol. 10, 1997.

[21] S. Williams, E. Reiter, and L. M. Osman, "Experiments with discourse-level choices and readability," in *In Proceedings of the European Natural Language Generation Workshop (ENLG) and 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL03)*, 2003, pp. 127–134.

[22] R. Barzilay and N. Elhadad, "Sentence alignment for monolingual comparable corpora," in *In Proceedings of the 2003 conference on Empirical methods in natural language processing*, 2003, pp. 25–32.

[23] A. Fujita, "Automatic generation of syntactically well-formed and semantically appropriate paraphrases," Ph.D. dissertation, Graduate School of Information Science, Nara Institute of Science and Technology (NAIST), 2005.

[24] R. Bhagat, "Learning paraphrases from text," Ph.D. dissertation, University of Southerm California, Los Angeles, CA, 2009.

[25] C. Fellbaum, Ed., *WordNet: an electronic lexical database*. MIT Press, 1998.

[26] D. Klein and C. D. Manning, "Fast exact inference with a factored model for natural language parsing," in *In Advances in Neural Information Processing Systems 15*, 2003, pp. 3–10.

[27] D. Lin, "Dependency-based evaluation of MINIPAR," in *Proc. of Workshop on the Evaluation of Parsing Systems*, Granada, Spain, 1998.

[28] Inclusion Europe Association, "Inclusion europe," *http://www.inclusion-europe.org* [7/2/2011], 1998.