Supporting the Adaptation of Texts for Poor Literacy Readers: a Text Simplification Editor for Brazilian Portuguese

Article · June 2009

CITATIONS
50

READS
76

6 authors, including:

Erick Galani Maziero
University of São Paulo
30 PUBLICATIONS **344** CITATIONS
SEE PROFILE

Caroline Gasperin
SwiftKey
37 PUBLICATIONS **603** CITATIONS
SEE PROFILE

Lucia Specia
The University of Sheffield
262 PUBLICATIONS **5,168** CITATIONS
SEE PROFILE

Sandra M. Aluisio
University of São Paulo
152 PUBLICATIONS **1,272** CITATIONS
SEE PROFILE

Supporting the Adaptation of Texts for Poor Literacy Readers: a Text Simplification Editor for Brazilian Portuguese

Arnaldo Candido Jr., Erick Maziero, Caroline Gasperin, Thiago A. S. Pardo, Lucia Specia, and Sandra M. Aluisio

Center of Computational Linguistics (NILC) / Department of Computer Sciences, University of São Paulo
Av. Trabalhador São-Carlense, 400. 13560-970 - São Carlos/SP, Brazil
arnaldoc@icmc.usp.br, egmaziero@gmail.com, {cgasperin,taspardo,lspecia,sandra}@icmc.usp.br

Abstract

In this paper we investigate the task of text simplification for Brazilian Portuguese. Our purpose is three-fold: to introduce a simplification tool for such language and its underlying development methodology, to present an on-line authoring system of simplified text based on the previous tool, and finally to discuss the potentialities of such technology for education. The resources and tools we present are new for Portuguese and innovative in many aspects with respect to previous initiatives for other languages.

1 Introduction

In Brazil, according to the index used to measure the literacy level of the population (INAF - National Indicator of Functional Literacy), a vast number of people belong to the so called rudimentary and basic literacy levels. These people are only able to find explicit information in short texts (rudimentary level) or process slightly longer texts and make simple inferences (basic level). INAF reports that 68% of the 30.6 million Brazilians between 15 and 64 years who have studied up to 4 years remain at the rudimentary literacy level, and 75% of the 31.1 million who studied up to 8 years remain at the rudimentary or basic levels.

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 is a part of (Snow, 2002). 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. When these factors do not match the reader's knowledge and experience, the text becomes too complex for the comprehension to occur. In this paper we will focus on the text and the aspects of it that make reading difficult or easy. One solution to ease the syntactic structure of a text is via Text Simplification (TS) facilities.

TS aims to maximize the comprehension of written texts through the simplification of their linguistic structure. This may involve simplifying lexical and syntactic phenomena, by substituting words that are only understood by a few people with words that are more usual, and by breaking down and changing the syntactic structure of the sentence, respectively. As a result, it is expected that the text can be more easily understood both by humans and computer systems (Mapleson, 2006; Siddharthan, 2003, Max, 2006). TS may also involve dropping parts or full sentences and adding some extra material to explain a difficult point. This is the case, for example, of the approach presented by Petersen and Ostendorf (2007), in which abridged versions of articles are used in adult literacy learning.

It has already been shown that long sentences, conjoined sentences, embedded clauses, passives, non-canonical word order, and use of low-frequency words, among other things, increase text complexity for language-impaired readers (Siddharthan, 2002; Klebanov et al., 2004; Devlin and Unthank, 2006). The Plain English initiative makes available guidelines to make texts easier to comprehend: the *Plain Language*[1]. In principle, its recommendations can be applied to any language. Although some of them are directly useful for TS systems (e.g., subject-verb-object order and active voice), others are difficult to specify (e.g., how simple each syntactic construction is and which words are simple).

In this paper we present the results of a study of syntactic simplification for Brazilian Portuguese (BP) and a rule-based syntactic simplification system for this language that was developed based on this study – the first of this kind for BP. We also present an on-line authoring tool for creating simplified texts. One possible application of this tool is to help teachers to produce instructional texts

[1] http://www.plainlanguage.gov

to be used in classrooms. The study is part of the PorSimples project[2] (Simplification of Portuguese Text for Digital Inclusion and Accessibility), which aims at producing text simplification tools for promoting digital inclusion and accessibility for people with different levels of literacy, and possibly other kinds of reading disabilities.

This paper is organized as follows. In Section 2 we present related approaches for text simplification with educational purposes. In Section 3 we describe the proposed approach for syntactic simplification, which is used within an authoring tool described in Section 4. In Section 5 we discuss possible uses of text simplification for educational purposes.

2 Related work

Burstein (2009) presents an NLP-based application for educational purposes, named Text Adaptor, which resembles our authoring tool. It includes complex sentence highlighting, text elaboration (word substitutions by easier ones), text summarization and translation. The system does not perform syntactic simplification, but simply suggests, using a shallow parser, that some sentences might be too complex. Specific hints on the actual source of complexity are not provided.

Petersen (2007) addresses 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 (Siddharthan, 2003) is used to split the selected sentences. In our approach, we do not drop sentences, since we believe that all the content must be kept in the text.

Siddharthan 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; it has not targeted any particular application. The system treats apposition, relative clauses, coordination and subordination. Our method, on the other hand, relies on deep parsing (Bick, 2000). We treat the same phenomena as Siddharthan, but also deal with Subject-Verb-Object ordering (in Portuguese sentences can be written in different orders) and passive to active voice conversion. Siddharthan's system deals with non-finite clauses which are not handled by our system at this stage.

Lal and Ruger's (2002) created a bayesian summarizer with a built-in lexical simplification module, based on WordNet and MRC psycho-linguistic database[3]. The system focuses on schoolchildren and provides background information about people and locations in the text, which are retrieved from databases. Our rule-based simplification system only replaces discourse markers for more common ones using lexical resources built in our project, instead of inserting additional information in the text.

Max (2005, 2006) 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. The resulting suggested simplifications are ranked by a score of syntactic complexity and potential change of meaning. The writer then chooses their preferred simplification. This system ensures accurate output, but requires human intervention at every step. Our system, on the other hand, is autonomous, even though the user is able to undo any undesirable simplification or to choose alternative simplifications. These alternative simplifications may be produced in two cases: i) to compose a new subject in simplifications involving relatives and appositions and ii) to choose among one of the coordinate or subordinate simplifications when there is ambiguity regarding to conjunctions.

Inui et al. (2003) proposes a rule-based system for text simplification aimed at deaf people. The authors create readability assessments based on questionnaires answered by teachers about the deaf. With approximately one thousand manually created rules, the authors generate several paraphrases for each sentence and train a classifier to select the simpler ones. Promising results are obtained, although different types of errors on the paraphrase generation are encountered, such as problems with verb conjugation and regency. In our work we produce alternative simplifications only in the two cases explained above.

[2] http://caravelas.icmc.usp.br/wiki/index.php/Principal

[3] http://www.psych.rl.ac.uk/

Caseli et al. (2009) developed an annotation editor to support the building of parallel corpora of original and simplified texts in Brazilian Portuguese. The tool was used to build a corpus of simplified texts aimed at people with rudimentary and basic literacy levels. We have used the parallel corpus to evaluate our rule-based simplification system. The on-line authoring system presented in this paper evolved from this annotation editor.

There are also commercial systems like Simplus[4] and StyleWriter[5], which aim to support Plain English writing.

3 A rule-based syntactic simplification system

Our text simplification system comprises seven operations (see Sections 3.1 and 3.2), which are applied to a text in order to make its syntactic structure simpler. These operations are applied sentence by sentence, following the 3-stage architecture proposed by Siddharthan (2002), which includes stages of analysis, transformation and regeneration. In Siddharthan's work, the analysis stage performs the necessary linguistic analyses of the input sentences, such as POS tagging and chunking; the transformation stage applies simplification rules, producing simplified versions of the sentences; the regeneration stage performs operations on the simplified sentences to make them readable, like referring expressions generation, cue words rearrangement, and sentence ordering. Differently from such architecture, currently our regeneration stage only includes the treatment of cue words and a surface forms (GSF) generator, which is used to adjust the verb conjugation and regency after some simplification operations.

As a single sentence may contain more than one complex linguistic phenomenon, simplification operations are applied in cascade to a sentence, as described in what follows.

3.1 Simplification cases and operations

As result of a study on which linguistic phenomena make BP text complex to read and how these phenomena could be simplified, we elaborated a manual of BP syntactic simplification (Aluisio et al., 2008). The rule-based text simplification system developed here is based on the specifications in this manual. According to this manual, simplification operations should be applied when any of the 22 linguistic phenomena presented in Table 1 is detected.

The possible operations suggested to be applied in order to simplify these phenomena are: (a) split the sentence, (b) change a discourse marker by a simpler and/or more frequent one (the indication is to avoid the ambiguous ones), (c) change passive to active voice, (d) invert the order of the clauses, (e) convert to subject-verb-object ordering, (f) change topicalization and detopicalization of adverbial phrases and (g) non-simplification.

Table 1 shows the list of all simplification phenomena covered by our manual, the clues used to identify the phenomena, the simplification operations that should be applied in each case, the expected order of clauses in the resulting sentence, and the cue phrases (translated here from Portuguese) used to replace complex discourse markers or to glue two sentences. In column 2, we consider the following clues: syntactic information (S), punctuation (P), and lexicalized clues, such as conjunctions (Cj), relative pronouns (Pr) and discourse markers (M), and semantic information (Sm, and NE for named entities).

3.2 Identifying simplification cases and applying simplification rules

Each sentence is parsed in order to identify cases for simplification. We use parser PALAVRAS (Bick, 2000) for Portuguese. This parser provides lexical information (morphology, lemma, part-of-speech, and semantic information) and the syntactic trees for each sentence. For some operations, surface information (such as punctuation or lexicalized cue phrases) is used to identify the simplification cases, as well as to assist simplification process. For example, to detect and simplify subjective non-restrictive relative clauses (where the relative pronoun is the subject of the relative clause), the following steps are performed:

1. The presence of a relative pronoun is verified.
2. Punctuation is verified in order to distinguish it from restrictive relative clauses: check if the pronoun occurs after a comma or semicolon.
3. Based on the position of the pronoun, the next punctuation symbol is searched to define the boundaries of the relative clause.

[4] http://www.linguatechnologies.com/english/home.html
[5] http://www.editorsoftware.com/writing-software

4. The first part of the simplified text is generated, consisting of the original sentence without the embedded relative clause.
5. The noun phrase in the original sentence to which the relative clause refers is identified.
6. A second simplified sentence is generated, consisting of the noun phrase (as subject) and the relative clause (without the pronoun).

The identification of the phenomena and the application of the operations are prone to errors though. Some of the clues that indicate the occurrence of the phenomena may be ambiguous. For example, some of the discourse markers that are used to identify subordinate clauses can indicate more than one type of these: for instance, "como" (in English "like", "how" or "as") can indicate reason, conformative or concessive subordinate clauses. Since there is no other clue that can help us disambiguate among those, we always select the case that occurs more frequently according to a corpus study of discourse markers and the rhetoric relations that they entitle (Pardo and Nunes, 2008). However, we can also treat all cases and let the user decide the simplifications that is most appropriate.

Phenomenon	***Clues***	***Op***	***Clause Order***	***Cue phrase***	***Comments***
1.Passive voice	S	c			Verb may have to be adapted
2.Embedded appositive	S	a	Original/ App.		Appositive: Subject is the head of original + to be in present tense + apposition
3.Asyndetic coordinate clause	S	a	Keep order		New sentences: Subjects are the head of the original subject
4.Additive coordinate clause	S, Cj	a	Keep order	Keep marker	Marker appears in the beginning of the new sentence
5.Adversative coordinate clause	M	a, b	Keep order	*But*	
6.Correlated coordinate clause	M	a, b	Keep order	*Also*	Original markers disappear
7.Result coordinate clause	S, M	a, b	Keep order	*As a result*	
8.Reason coordinate clause	S, M	a, b	Keep order	*This happens because*	May need some changes in verb
9.Reason subordinate clause	M	a, b, d	Sub/Main	*With this*	To keep the ordering cause, result
10.Comparative subordinate clause	M	a, b	Main/Sub	*Also*	Rule for *such ... as, so ... as* markers
	M	g			Rule for the other markers or short sentences
11.Concessive subordinate clause	M	a, b, d	Sub/Main	*But*	"Clause 1 *although* clause 2" is changed to "Clause 2. *But* clause 1"
	M	a, b	Main/Sub	*This happens even if*	Rule for hypothetical sentences
12.Conditional subordinate clause	S, M	d	Sub/Main		Pervasive use in simple accounts
13. Result subordinate clause	M	a, b	Main/Sub	*Thus*	May need some changes in verb
14.Final/Purpose subordinate clause	S, M	a, b	Main/Sub	*The goal is*	
15.Confirmative subordinate clause	M	a, b, d	Sub/Main	*Confirms that*	May need some changes in verb
16.Time subordinate clause	M	a	Sub/Main		May need some changes in verb
	M	a, b		*Then*	Rule for markers: after that, as soon as
17. Proportional Subordinate Clause	M	g			
18. Non-finite subordinate clause	S	g			
19.Non-restrictive relative clause	S, P, Pr	a	Original/ Relative		Relative: Subject is the head of original + relative (subjective relative clause)
20.Restrictive relative clause	S, Pr	a	Relative/ Original		Relative: Subject is the head of original + relative (subjective relative clause)
21.Non Subject-Verb-Object order	S	e			Rewrite in Subject-Verb-Object order
22. Adverbial phrases in theme position	S, NE, Sm	f	In study		In study

Table 1: Cases, operations, order and cue phrases

Every phenomenon has one or more simplification steps associated with it, which are applied to perform the simplification operations. Below we detail each operation and discuss the challenges involved and our current limitations in their implementing.

a) Splitting the sentence - This operation is the most frequent one. It requires finding the split point

in the original sentence (such as the boundaries of relative clauses and appositions, the position of coordinate or subordinate conjunctions) and the creation of a new sentence, whose subject corresponds to the replication of a noun phrase in the original sentence. This operation increases the text length, but decreases the length of the sentences. With the duplication of the term from the original sentence (as subject of the new sentence), the resulting text contains redundant information, but it is very helpful for people at the rudimentary literacy level.

When splitting sentences due to the presence of apposition, we need to choose the element in the original sentence to which it is referring, so that this element can be the subject of the new sentence. At the moment we analyze all NPs that precede the apposition and check for gender and number agreement. If more than one candidate passes the agreement test, we choose the closest one among these; if none does, we choose the closest among all candidates. In both cases we can also pass the decision on to the user, which we do in our authoring tool described in Section 4.

For treating relative clauses we have the same problem as for apposition (finding the NP to which the relative clause is anchored) and an additional one: we need to choose if the referent found should be considered the subject or the object of the new sentence. Currently, the parser indicates the syntactic function of the relative pronoun and that serves as a clue.

b) Changing discourse marker - In most cases of subordination and coordination, discourse markers are replaced by most commonly used ones, which are more easily understood. The selection of discourse markers to be replaced and the choice of new markers (shown in Table 1, col. 4) are done based on the study of Pardo and Nunes (2008).

c) Transformation to active voice - Clauses in the passive voice are turned into active voice, with the reordering of the elements in the clause and the modification of the tense and form of the verb. Any other phrases attached to the object of the original sentence have to be carried with it when it moves to the subject position, since the voice changing operation is the first to be performed. For instance, the sentence:

> "More than 20 people have been bitten by gold piranhas (Serrasalmus Spilopleura), which live in the waters of the Sanchuri dam, next to the BR-720 highway, 40 km from the city."

is simplified to:

> "Gold piranhas (Serrasalmus Spilopleura), which live in the waters of the Sanchuri dam, next to the BR-720 highway, 40 km from the city, have bitten more than 20 people."

After simplification of the relative clause and apposition, the final sentence is:

> "Gold piranhas have bitten more than 20 people. Gold piranhas live in the waters of the Sanchuri dam, next to the BR-720 highway, 40 km from the city. Gold piranhas are Serrasalmus Spilopleura."

d) Inversion of clause ordering - This operation was primarily designed to handle subordinate clauses, by moving the main clause to the beginning of the sentence, in order to help the reader processing it on their working memory (Graesser et al., 2004). Each of the subordination cases has a more appropriate order for main and subordinate clauses (as shown in Table 1, col. 3), so that "independent" information is placed before the information that depends on it. In the case of concessive subordinate clauses, for example, the subordinate clause is placed before the main clause. This gives the sentence a logical order of the expressed ideas. See the example below, in which there is also a change of discourse marker and sentence splitting, all operations assigned to concessive subordinate clauses:

> "The building hosting the Brazilian Consulate was also evacuated, although the diplomats have obtained permission to carry on working."

Its simplified version becomes:

> "The diplomats have obtained permission to carry on working. But the building hosting the Brazilian Consulate was also evacuated."

e) Subject-Verb-Object ordering - If a sentence is not in the form of subject-verb-object, it should be rearranged. This operation is based only on information from the syntactic parser. The example below shows a case in which the subject is after the verb (translated literally from Portuguese, preserving the order of the elements):

> "On the 9th of November of 1989, fell the wall that for almost three decades divided Germany."

Its simplified version is:

> "On the 9th of November of 1989, the wall that for almost three decades divided Germany fell."

Currently the only case we are treating is the non-canonical order Verb-Object-Subject. We plan to treat other non-canonical orderings in the near future. Besides that, we still have to define how to deal with elliptic subjects and impersonal verbs (which in Portuguese do not require a subject).

When performing this operation and the previous one, a generator of surface forms (GSF) is used to adjust the verb conjugation and regency. The GSF is compiled from the Apertium morphological dictionaries enhanced with the entries of Unitex-BP (Muniz et al., 2005), with an extra processing to map the tags of the parser to those existing in morphological dictionaries (Caseli et al., 2007) to obtain an adjusted verb in the modified sentence.

f) Topicalization and detopicalization - This operation is used to topicalize or detopicalize an adverbial phrase. We have not implemented this operation yet, but have observed that moving adverbial phrases to the end or to the front of sentences can make them simpler in some cases. For instance, the sentence in the last example would become:

"The wall that for almost three decades divided Germany fell on the 9th of November of 1989."

We are still investigating how this operation could be applied, that is, which situations require (de)topicalization.

3.3 The cascaded application of the rules

As previously mentioned, one sentence may contain several phenomena that could be simplified, and we established the order in which they are treated. The first phenomenon to be treated is passive voice. Secondly, embedded appositive clauses are resolved, since they are easy to simplify and less prone to errors. Thirdly, subordinate, non-restrictive and restrictive relative clauses are treated, and only then the coordinate clauses are dealt with.

As the rules were designed to treat each case individually, it is necessary to apply the operations in cascade, in order to complete the simplification process for each sentence. At each iteration, we (1) verify the phenomenon to be simplified following the standard order indicated above; (2) when a phenomenon is identified, its simplification is executed; and (3) the resulting simplified sentence goes through a new iteration. This process continues until there are no more phenomena. The cascade nature of the process is crucial because the simplified sentence presents a new syntactic structure and needs to be reparsed, so that the further simplification operations can be properly applied. However, this process consumes time and is considered the bottleneck of the system.

3.4 Simplification evaluation

We have so far evaluated the capacity of our rule-based simplifier to identify the phenomena present in each sentence, and to recommend the correct simplification operation. We compared the operations recommended by the system with the ones performed manually by an annotator in a corpus of 104 news articles from the Zero Hora newspaper, which can be seen in our Portal of Parallel Corpora of Simplified Texts[6]. Table 2 presents the number of occurrences of each simplification operation in this corpus.

Simplification Operations	***# Sentences***
Non-simplification	2638
Subject-verb-object ordering	44
Transformation to active voice	154
Inversion of clause ordering	265
Splitting sentences	1103

Table 2. Statistics on the simplification operations

The performance of the system for this task is presented in Table 3 in terms of precision, recall, and F-measure for each simplification operation.

Operation	***P***	***R***	***F***
Splitting sentences	64.07	82.63	72.17
Inversion of clause ordering	15.40	18.91	16.97
Transformation to active voice	44.29	44.00	44.14
Subject-verb-object ordering	1.12	4.65	1.81
ALL	51.64	65.19	57.62
Non-simplification	64.69	53.58	58.61

Table 3. Performance on defining simplification operations according to syntactic phenomena

These results are preliminary, since we are still refining our rules. Most of the recall errors on the inversion of clause ordering are due to the absence of a few discourse markers in the list of markers that we use to identify such cases. The majority of recall errors on sentence splitting are due to mistakes on the output of the syntactic parser and to the number of ordering cases considered and implemented so far. The poor performance for subject-verb-object ordering, despite suffering from mistakes of the parser, indicates that our rules for this operation need to be refined. The same applies to inversion of clause ordering.

We did not report performance scores related to the "changing discourse marker" operation because in our evaluation corpus this operation is merged with other types of lexical substitution. However, in

[6] http://caravelas.icmc.usp.br/portal/index.php

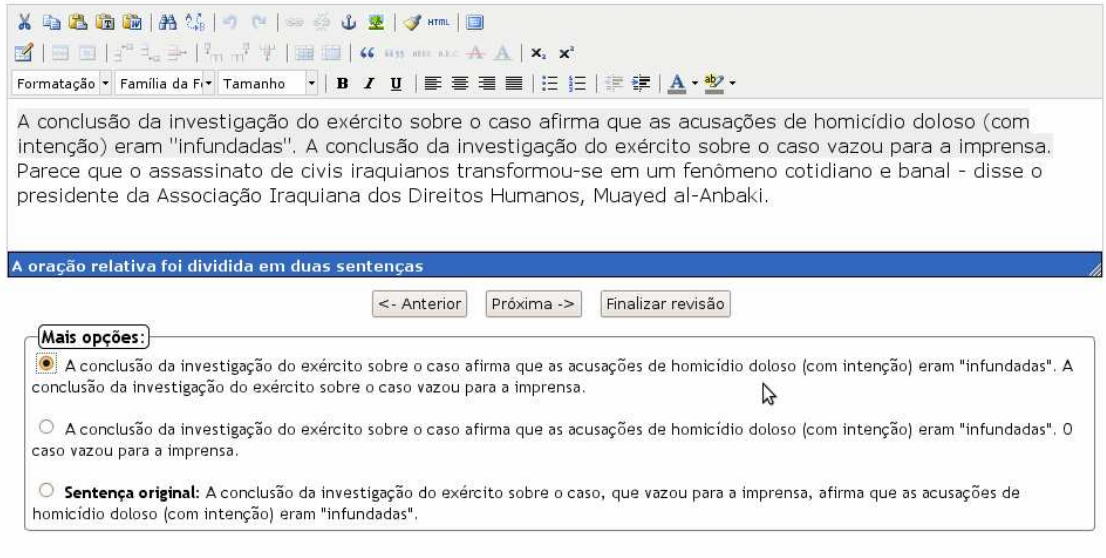

Figure 1: Interface of the *Simplifica* system

order to assess if the sentences were correctly simplified, it is necessary to do a manual evaluation, since it is not possible to automatically compare the output of the rule-based simplifier with the annotated corpus, as the sentences in the corpus have gone through operations that are not performed by the simplifier (such as lexical substitution). We are in the process of performing such manual evaluation.

4 *Simplifica* editor: supporting authors

We developed *Simplifica*[7] (Figure 1), an authoring system to help writers to produce simplified texts. It employs the simplification technology described in the previous section. It is a web-based WYSIWYG editor, based on TinyMCE web editor[8].

The user inputs a text in the editor, customizes the simplification settings where one or more simplifications can be chosen to be applied in the text and click on the "simplify" button. This triggers the syntactic simplification system, which returns an XML file containing the resulting text and tags indicating the performed simplification operations. After that, the simplified version of the text is shown to the user, and he/she can revise the automatic simplification.

4.1 The XML representation of simplification operations

Our simplification system generates an XML file describing all simplification operations applied to a text. This file can be easily parsed using standard XML parsers. Table 5 presents the XML annotation to the "gold piranhas" example in Section 3.2.

```
<simplification type="passive">
  <simplification type="appositive">
    <simplification type="relative">
        Gold piranhas have bitten more than 20 people. Gold
        piranhas live in the waters of the Sanchuri dam, next to
        the BR-720 highway, 40 km from the city.
    </simplification>
    Gold piranhas are Serrasalmus Spilopleura.
  </simplification>
</simplification>
```

Table 5. XML representation of a simplified text

In our annotation, each sentence receives a <simplification> tag which describes the simplified phenomena (if any); sentences that did not need simplification are indicated with a <simplification type="no"> tag. The other simplification types refer to the eighteen simplification cases presented in Table 1. Nested tags indicate multiple operations applied to the same sentence.

4.2 Revising the automatic simplification

Once the automatic simplification is done, a review screen shows the user the simplified text so that he/she can visualize all the modifications applied and approve or reject them, or select alternative simplifications. Figure 1 shows the reviewing screen and a message related to the simplification performed below the text simplified.

The user can revise simplified sentences one at a time; the selected sentence is automatically highlighted. The user can accept or reject a

[7] http://www.nilc.icmc.usp.br/porsimples/simplifica/
[8] http://tinymce.moxiecode.com/

simplified sentence using the buttons below the text. In the beginning of the screen "Mais opções", alternative simplifications for the sentence are shown: this facility gives the user the possibility to resolve cases known to be ambiguous (as detailed in Sections 2 and 3.2) for which the automatic simplification may have made a mistake. In the bottom of the same screen we can see the original sentence ("Sentença original") to which the highlighted sentence refers.

For the example in Figure 1, the tool presents alternative simplifications containing different subjects, since selecting the correct noun phrase to which an appositive clause was originally linked (which becomes the subject of the new sentence) based on gender and number information was not possible.

At the end of the process, the user returns to the initial screen and can freely continue editing the text or adding new information to it.

5 Text Simplification for education

Text simplification can be used in several applications. Journalists can use it to write simple and straightforward news texts. Government agencies can create more accessible texts to a large number of people. Authors of manuals and technical documents can also benefit from the simplification technology. Simplification techniques can also be used in an educational setting, for example, by a teacher who is creating simplified texts to students. Classic literature books, for example, can be quite hard even to experienced readers. Some genres of texts already have simplified versions, even though the simplification level can be inadequate to a specific target audience. For instance, 3rd and 7th grade students have distinct comprehension levels.

In our approach, the number and type of simplification operations applied to sentences determine its appropriateness to a given literacy level, allowing the creation of multiple versions of the same text, with different levels of complexity, targeting special student needs.

The *Simplifica* editor allows the teacher to adopt any particular texts to be used in the class, for example, the teacher may wish to talk about current news events with his/her students, which would not be available via any repository of simplified texts. The teacher can customize the text generating process and gradually increase the text complexity as his/her students comprehension skills evolve. The use of the editor also helps the teacher to develop a special awareness of the language, which can improve his/her interaction with the students.

Students can also use the system whenever they have difficulties to understand a text given in the classroom. After a student reads the simplified text, the reading of the original text becomes easier, as a result of the comprehension of the simplified text. In this scenario, reading the original text can also help the students to learn new and more complex words and syntactic structures, which would be harder for them without reading of the simplified text.

6 Conclusions

The potentialities of text simplification systems for education are evident. For students, it is a first step for more effective learning. Under another perspective, given the Brazilian population literacy levels, we consider text simplification a necessity. For poor literacy people, we see text simplification as a first step towards social inclusion, facilitating and developing reading and writing skills for people to interact in society. The social impact of text simplification is undeniable.

In terms of language technology, we not only introduced simplification tools in this paper, but also investigated which linguistic phenomena should be simplified and how to simplify them. We also developed a representation schema and designed an on-line authoring system. Although some aspects of the research are language dependent, most of what we propose may be adapted to other languages.

Next steps in this research include practical applications of such technology and the measurement of its impact for both education and social inclusion.

Acknowledgments

We thank the Brazilian Science Foundation FAPESP and Microsoft Research for financial support.

References

Aluísio, S.M., Specia, L., Pardo, T.A.S., Maziero, E.G., Fortes, R. 2008. Towards Brazilian Portuguese Automatic Text Simplification Systems. In the *Proceedings of the 8th ACM Symposium on Document Engineering*, pp. 240-248.

Bick, E. 2000. *The parsing system "Palavras"*:

Automatic grammatical analysis of Portuguese in a constraint grammar framework. PhD Thesis University of Århus, Denmark.

Burstein, J. 2009. Opportunities for Natural Language Processing Research in Education. In the *Proceedings of CICLing*, pp. 6-27.

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

Caseli, H.M.; Nunes, M.G.V.; Forcada, M.L. 2008. Automatic induction of bilingual resources from aligned parallel corpora: application to shallow-transfer machine translation. *Machine Translation*, V. 1, p. 227-245.

Devlin, S., Unthank, G. 2006. Helping aphasic people process online information. In the *Proceedings of the ACM SIGACCESS Conference on Computers and Accessibility*, pp. 225-226.

Graesser, A., McNamara, D. S., Louwerse, M., Cai, Z. 2004. Coh-Metrix: Analysis of text on cohesion and language. *Behavioral Research Methods, Instruments, and Computers*, V. 36, pp. 193-202.

Inui, K., Fujita, A., Takahashi, T., Iida, R., Iwakura, T. 2003. Text Simplification for Reading Assistance: A Project Note. In the *Proceedings of the Second International Workshop on Paraphrasing*, 9 -16.

Klebanov, B., Knight, K., Marcu, D. 2004. Text Simplification for Information-Seeking Applications. *On the Move to Meaningful Internet Systems*. LNCS, V.. 3290, pp. 735-747.

Lal, P., Ruger, S. 2002. Extract-based summarization with simplification. In the *Proceedings of DUC*.

Mapleson, D.L. 2006. *Post-Grammatical Processing for Discourse Segmentation*. PhD Thesis. School of Computing Sciences, University of East Anglia, Norwich.

Max, A. 2005. Simplification interactive pour la production de textes adapt´es aux personnes souffrant de troubles de la compr´ehension. In the *Proceedings of Traitement Automatique des Langues Naturelles (TALN)*.

Max, A. 2006. Writing for language-impaired readers. In the *Proceedings of CICLing*, pp. 567-570.

Muniz, M.C., Laporte, E. Nunes, M.G.V. 2005. UNITEX-PB, a set of flexible language resources for Brazilian Portuguese. In *Anais do III Workshop em Tecnologia da Informação e da Linguagem Humana*, V. 1, pp. 1-10.

Pardo, T.A.S. and Nunes, M.G.V. 2008. On the Development and Evaluation of a Brazilian Portuguese Discourse Parser. *Journal of Theoretical and Applied Computing*, V. 15, N. 2, pp. 43-64.

Petersen, S.E. 2007. *Natural Language Processing Tools for Reading Level Assessment and Text Simplification for Bilingual Education*. PhD Thesis, University of Washington.

Petersen, S.E. and Ostendorf, M. 2007. Text Simplification for Language Learners: A Corpus Analysis. In the *Proceedings of the Speech and Language Technology for Education Workshop*, pp. 69-72.

Specia, L., Aluísio, S.M., Pardo, T.A.S. 2008. *Manual de simplificação sintática para o português*. Technical Report NILC-TR-08-06, NILC.

Siddharthan, A. 2002. An Architecture for a Text Simplification System. In the *Proceedings of the Language Engineering Conference*, pp. 64-71.

Siddharthan, A. 2003. *Syntactic Simplification and Text Cohesion*. PhD Thesis. University of Cambridge.

Snow, C. 2002. *Reading for understanding: Toward an R&D program in reading comprehension*. Santa Monica, CA.