Building a Brazilian Portuguese parallel corpus of original and simplified texts

Helena M. Caseli[1], Tiago F. Pereira[1], Lucia Specia[1], Thiago A. S. Pardo[1], Caroline Gasperin[1], and Sandra M. Aluisio[1]

[1]Center of Computational Linguistics (NILC)/ Department of Computer Sciences, University of São Paulo, Av. Trabalhador São-Carlense, 400. 13560-970 - São Carlos/SP, Brazil
helenacaseli@dc.ufscar.br, tiagofrepereira@yahoo.com.br, lspecia@icmc.usp.br, taspardo@icmc.usp.br, cgasperin@icmc.usp.br, sandra@icmc.usp.br

Abstract. In this paper we address the problem of building the necessary tools and resources for performing Brazilian Portuguese text simplification. We describe our efforts on the design and development of: (a) a XCES-based annotation schema, (b) an annotation edition tool, and (c) a portal to access parallel corpora of original-simplified texts. These contributions were intended to (i) allow the creation and public release of a corpus of original and simplified texts with two different versions of simplification (called here *natural* and *strong*), targeting two levels of functional illiteracy and (ii) register simplification decisions during the creation of such corpus. We also provide an analysis of the first corpus created using the resources presented here: 104 newspaper texts and their simplified versions, produced by an expert in text simplification.

Keywords: Text Simplification, Brazilian Portuguese, annotation standards, annotation edition tool.

1 Introduction

In Brazil, "letramento" (literacy) is the term used to designate people's ability to use written language to obtain and record information, express themselves, plan and learn continuously [1]. 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 able to find explicit information in short texts (rudimentary level) and also process slightly longer texts and make simple inferences (basic level).

The PorSimples project (*Simplificação Textual do Português para Inclusão e Acessibilidade Digital*)[1] aims at producing text simplification tools for promoting digital inclusion and accessibility for people with such levels of literacy, and possibly other kinds of reading disabilities. More specifically, the goal is to help these readers

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

to process documents available on the web. Additionally, it could help children learning to read texts of different genres or adults being alphabetized. Two tools are envisioned: (1) a browser plugin, which automatically simplifies texts on the web for the end-user, and (2) an authoring tool, which supports authors in the process of producing simple texts. The focus is on texts published in government sites or by relevant news agencies, both expected to be of importance to a large audience with various literacy levels. The language of the texts is Brazilian Portuguese, for which there are no text simplification systems, to the best of our knowledge.

The project follows three main text processing strategies to produce simplified texts: (i) text summarization, (ii) highlighting of the text structure/organization, named entities and verb-argument structure, aiming to provide visual and explanatory information about important concepts appearing in the text, and mainly (iii) text simplification itself, which includes operations at the lexical, syntactic and discourse levels. The simplification operations proposed in the project aim to preserve most of the information in the input text, and thus the deletion of a sentence or parts of it was rarely adopted. For that reason, summarization techniques play an important role.

Text simplification has been exploited in other languages for helping poor literacy readers [2], [3] and [4] and special kinds of readers such as aphasics [5]. It has also been used for improving the accuracy of other Natural Language Processing (NLP) tasks [6] and [7], like parsing. One important step towards building text simplification tools is the analysis and comparison of general-use, non-simplified texts, with their corresponding simplified versions, that is, a parallel corpus of original-simplified texts. This allows investigating which kinds of changes should be applied, what resources are necessary to allow them, and how to evaluate the simplification task. Moreover, such a corpus can be directly used with statistical techniques to learn simplification rules.

A corpus of original and manually simplified sentences has been created for English but it is no longer available [8]. However, such a resource does not contain any explicit information about how and why the simplifications were performed, and therefore only limited learning from this corpus is possible. Two other studies have used parallel aligned corpus of original and simplified English texts. [9] uses parallel corpora of TV program transcripts and subtitles (documentaries and talk shows broadcasted by the BBC World Service) to automatically generate subtitles for hearing-impaired people. [10] uses a corpus of original news articles with corresponding abridged versions developed by Literacyworks[2] to aid teachers by automatically proposing ways to simplify texts.

Such parallel corpora of original and simplified texts do not exist for Portuguese. Moreover, given the differences between the two languages, a parallel corpus of English simplifications would not be appropriate. So, in the scope of the PorSimples project we have: (1) built a parallel corpus of original and simplified texts for Brazilian Portuguese, (2) developed a tool to assist human annotators in this inherently manual task — the Simplification Annotation Editor[3] — and (3) specified a new schema for representing the original-simplified information, based on the XCES

[2] http://literacynet.org/cnnsf/index cnnsf.html
[3] http://caravelas.icmc.usp.br/anotador/

standard[4]. The parallel corpora resulting from the simplification process can be queried in a public Portal of Parallel Corpora of Simplified Texts[5].

The Simplification Annotation Editor facilitates the manual simplification task, by guiding the annotator and providing the necessary linguistic resources, besides recording the simplification operations made by the annotator. Moreover, as a consequence, it guarantees the consistency of the annotated corpora. The annotation process, on the other hand, also helps our understanding of the simplification task which can bring improvements to the tool, making it more comprehensive and compact..

This paper is organized as follows. In Section 2 we present the background and technologies related to this work. In Section 3 we describe the Simplification Annotation Editor and the Portal of Parallel Corpora of Simplified Texts, which shows all the simplification decisions taken in the annotation process for a given corpus. We also describe our XCES-based schema proposed to annotate simplification operations and present some statistics on a parallel corpus built using the Editor. In Section 4 we discuss some final remarks and present directions for future work.

2 Background and Related Work

2.1 Support Tools for Text Annotation and Simplification Editors

Text annotation is the process of adding new information to existing language data/corpora [11]. This is an inherently manual task, but it can be supported by tools. Some tools, such as GATE[6] and its several plugged-in systems, were developed to automatically annotate a corpus. MMAX (MultiModal Annotation in XML), another linguistic annotation tool, allows multi-level annotation of (potentially multi-modal) corpora [11]. Although very useful for several applications, the existing tools could not be used in for our purposes. GATE would require a system to be developed from scratch and MMAX is not able to specify the relations between different texts - the original and the simplified -, an essential piece of information in the text simplification annotation process.

There are also tools called *simplification editors*, such as SIMPLUS[7] and StyleWriter[8]. SIMPLUS is a generic tool for helping writing simplified (or controlled) English. Simplified English implies the use of limited vocabulary of Standard or Plain English words and restricted sentence structure. StyleWriter has also features to help users to write using Plain English. It guides the user on how to produce a well-written English text and also focus on simplifying and clarifying such text. Some simplification features present in these previous tools are included in our editor. However, instead of helping authors to write simple texts, currently, our editor is

[4] http://www.xml-ces.org
[5] http://caravelas.icmc.usp.br/portal/index.php
[6] http://gate.ac.uk/
[7] http://www.linguatechnologies.com
[8] http://www.editorsoftware.com/writing-software

intended to support the building of a parallel corpus of original-simplified texts to be used in corpus-driven approaches to text simplification. Therefore, besides the result of the simplification process, we need also to record the simplification operations that were performed. Other motivations for creating our own editor are that it is intended to be freely available to the research community and to evolve with the project, ultimately becoming a text simplification editor itself.

2.2 XCES

XCES is a corpus encoding standard in which the source documents are plain texts and all the annotations are stored in stand-off XML[9] documents [12]. The stand-off format for annotations is a graph representation in which the nodes are virtually placed between the characters in the plain text and the edges define regions between nodes, represented by XML annotations which are associated with feature structures [13]. For example, Figure 1 shows an excerpt of a stand-off annotation document containing the tokens of the Portuguese sentence in (snt_1). In this example, each *<struct>* element represents an edge in the graph and the values specified by the *from* and *to* attributes are the nodes in the source text document over which the edge spans. For example, the first token, "*Joni*" spans from node 270 (placed before character 'J') to node 274 (placed after character 'i') in the text document. The *<feat>* elements allow specifying any other relevant information about the element, such as its identifier and the actual word it represents.

(snt_1) *Joni Simões é proprietário de uma empresa da Capital que vende equipamentos de DVD. (Joni Simões owns a company in the capital which sells DVD devices).*

```
<struct type="token" from="270" to="274">
  <feat name="id" value="t47"/>
  <feat name="base" value="Joni"/>
</struct>
<struct type="token" from="275" to="281">
  <feat name="id" value="t48"/>
  <feat name="base" value="Simões"/>
</struct>
<struct type="token" from="282" to="283">
  <feat name="id" value="t49"/>
  <feat name="base" value="é"/>
</struct>
<struct type="token" from="284" to="296">
  <feat name="id" value="t50"/>
  <feat name="base" value="proprietário"/>
</struct>
```

Fig. 1. Excerpt of a stand-off XCES annotation document

XCES has been used in projects involving both only one language, e.g.: American National Corpus (ANC)[10] (English) and PLN-BR[11] (Brazilian Portuguese); and multiple languages as parallel data, e.g.: CroCo[12] (English-German) and Swedish-Turkish [14]. However, to our knowledge, PorSimples is the first project to use XCES to encode original-simplified parallel texts and also the actual simplification operations. Two annotation layers have been added to the traditional stand-off annotation layers, in order to store the information related to simplification.

[9] http://www.w3.org/XML/
[10] http://americannationalcorpus.org
[11] http://www.nilc.icmc.usp.br/plnbr
[12] http://fr46.uni-saarland.de/croco/index_en.html

In our XCES schema, each plain text document is related to at most other eight annotation documents, which contain the following information: (1) the header (specifies the origin of the document content and the stand-off annotation files), (2) the logical division (markup of the structure of the document), (3) the sentences (markup of the sentence boundaries), (4) the tokens, (5) the part-of-speech of the tokens, (6) the syntactic chunks (phrases), (7) the alignment between original and simplified sentences, and (8) the simplification operations performed to transform the original sentences into simplified sentences. The first five files follow the same formats of ANC and PLN-BR corpora. The sixth file is particularly important to build syntactic simplification systems both rule-based and statistical ones. The last two files also follow the XCES guidelines but were created specifically for this project (see Section 3.2).

2.3 The Use of Corpus for Text Simplification

Parallel corpora of original and simplified texts can be used for automatic text simplification considering: (1) the information obtained from the annotation process, and (2) the final result of this process (the actual annotated corpus). The first refers to the insights about the range of operations performed in order to simplify a text. These insights can guide the specification of a comprehensive and consistent set of simplification rules for rule-based simplification systems. The second refers to the several ways the parallel corpus can be used to design automatic text simplification systems by means of statistical or machine learning techniques.

[8] investigates the automatic induction of syntactic simplification rules from a parallel corpus. Syntactic correspondences are extracted and generalized into rules, for example, replacing words by variables. The work only covered isolating relative clauses and no evaluation was provided. [9] applies a case-based learning algorithm to a parallel corpus, focusing on the summarization of subtitles by the removal of elements and lexical substitution. A very low performance was reported and the system seems to make serious mistakes, such as removing the subject of the sentences. Both corpora developed in such investigations aim at the simplification of English texts. Details about the creation of these corpora are not discussed in the published materials, but since fewer simplification operations were covered, as compared to our set of operations, we believe that such a process was simpler. It appears that no tool was designed to help the annotators.

[3] and [10] present a detailed corpus analysis of original and manually simplified news articles aiming at learning how people simplify texts in order to develop better automatic tools. They focus on the features of sentences that are split and on position and redundancy information in decisions about which sentences to keep and which to drop. However, they did not develop a simplification system based on the outcome of the corpus analysis; instead they used the syntactic simplifier of [4].

We believe that with a well designed and appropriately annotated corpus of original-simplified texts, covering enough examples of the simplification operations aimed by the PorSimples project, we will be able to further investigate the learning techniques which can be applied (and most likely adapted) to this application.

3 Text Simplification Annotation in the PorSimples project

3.1 The Annotation Editor and the Portal of Parallel Corpora of Simplified Texts

As described in Section 1, readers with literacy at basic level may need different type of help from those with literacy at rudimentary level, and the same goes to children learning to read or people with cognitive disabilities. To attend the needs of people with different levels of literacy, we propose two subsets of simplifications called *natural* and *strong* simplifications. In our annotation tool, when performing a natural simplification, the annotator is free to choose which operations to use, among the ones available, and when to use them; there may be cases where the annotator decides not to simplify a sentence. Strong simplification, on the other hand, is driven by explicit rules from a manual of syntactic simplification also developed in the project [15] and [16], which state when and how to apply the simplification operations. Table 1 shows examples of an original text from an on-line Brazilian newspaper (translated here from Portuguese) in (a), its natural simplification in (b) and its strong simplification in (c). Clearly, the sentence in (b) can be further simplified if broken in shorter ones, as shown in (c). Although (c) may look less cohesive and somehow redundant, it can be useful for people with very low literacy levels [17].

Table 1. An example of an original text (a) and its simplified versions (b and c)

A	*In a press conference called to answer corruption charges during his term as Mayor of the city of Ribeirão Preto, Minister Antonio Palocci Filho (Treasury) said he made his position available, but with the recommendation of President Luiz Inácio Lula da Silva, would remain in government.*
B	*Minister Antonio Palocci (Treasury) said in a press conference that he will leave his position, although President Lula advised him to remain in the government.*
C	*Minister Antonio Palocci is the Treasury Minister. Antonio Palocci said in a press conference that he will leave his position. But he said that President Lula advised him to remain in the government.*

The Simplification Annotation Editor was used by the human annotator to create the parallel corpus following the 3-step architecture shown in Figure 2.

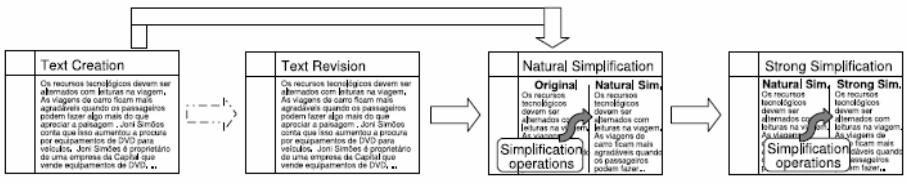

Fig. 2. Architecture of the Simplification Annotation Editor

In the first step, the source text (original version) is created (or simply opened from a file) and possibly revised. In the revision step, the human annotator may manually

correct punctuation and spelling mistakes. In the second step, natural simplifications are produced and logged, and from these the strong simplifications are generated (step3) (this sequence, first natural then strong, is not enforced in the Editor, that is, it allows strong simplifications from the original text as well). All the text versions (original, revised, natural and strong simplified) are stored in a database (DB).

To explain how the annotation is performed by a human using the Editor, consider the simplification example presented in Figure 3. This figure shows a screenshot of the Editor in the strong simplification step. As the numbers in Figure 3 show, the editor has three main areas: (1) the text being simplified, (2) the simplified version being produced, and (3) the log of simplification operations performed so far. In Figure 3, it is registered that the fourth original sentence, shown here in (snt_1) ("*Sentença: 4*") was divided in 2 sentences, as shown in snt_2 and snt_3).

(snt_2) *Joni Simões é proprietário de uma empresa da Capital (Joni Simões owns a company in the capital).*

(snt_3) *A empresa vende equipamentos de DVD (The company sells DVD devices).*

The simplification operations that can be applied encompass lexical and syntactic modifications and are performed for each original sentence separately. The syntactic operations, which are accessible via a pop-up menu, are the following: (1) non-simplification; (2) simple or (3) strong rewriting (as defined in [10]); (4) putting the sentence in its canonical order (subject-verb-object); (5) putting the sentence in the active voice; (6) inverting the clause ordering; (7) splitting or (8) joining sentences; (9) dropping the sentence or (10) dropping parts of the sentence. The lexical operations consist in replacing words found to be complex by simpler synonyms.

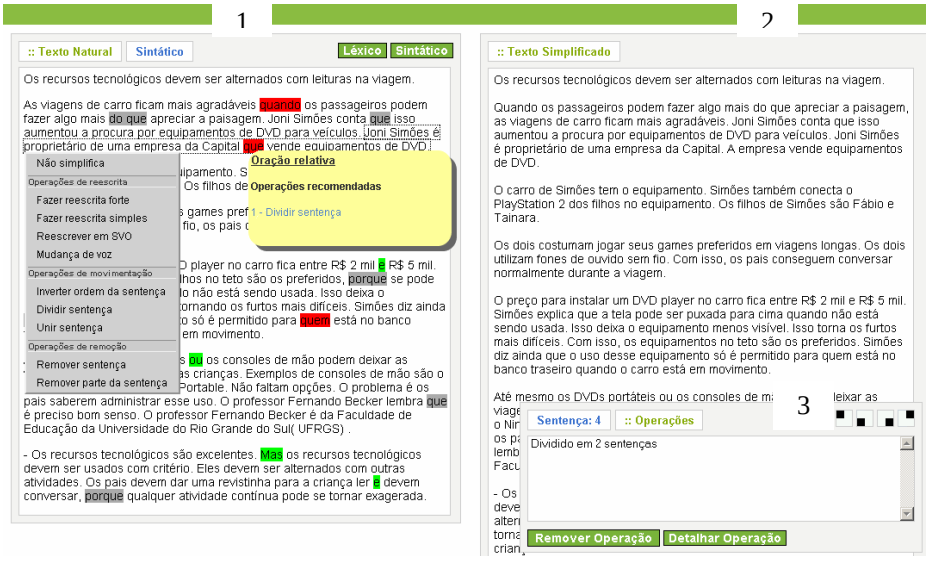

Fig. 3. Screenshot of the Simplification Annotation Editor (in the *Sintático* mode)

The Annotation Editor has two modes to assist the human annotator: the *Léxico* and the *Sintático* modes. In the *Léxico* mode, the editor proposes changes in words

and discourse markers by simpler and/or more frequent ones. The annotator decides whether to accept or not the suggestions to simplify the highlighted words. Lexical simplifications are performed based on two linguistic resources: (1) a list of simple words and (2) a list of discourse markers. The first list is composed of words supposed to be common to youngsters, extracted from [18], frequent words from news texts for children, and concrete words [19]. The discourse markers were extracted from [20]. The *Sintático* mode proposes the 10 previously mentioned syntactic operations based on syntactic information provided by a parser for Portuguese [21]. As an example, in Figure 3, the system recommends (in the recommendation box) splitting snt_1 ("*1- Dividir sentença*"), since it has a relative clause (introduced by the relative pronoun "*que*"). This operation can be either selected from the recommendation box or from the pop-up menu. When chosen, the operation is recorded (area (3) of Figure 3) and for each simplification operation it is possible to specify (in "*Detalhar operação*") what has been changed in the simplified version.

The resulting parallel corpus can be queried in the Portal of Parallel Corpora of Simplified Texts, which shows all the simplification operations performed. For example, one can recover all the original sentences that were split during simplification or see all the lexical substitution pairs composed of complex and simple words. The Portal also makes available the XCES annotation and the resources that were used, including the dictionaries of simple words and discourse markers. It allows searching the corpus for the original and simplified texts, the alignment between such texts, the syntactical constructions that were considered in the project, and the actual texts that underwent the simplification operations.

3.2 The XCES Output

The output of the simplification process consists of eight XCES files, as described in Section 2.2.

a
```
<struct type="opr">
   <feat name="id" value="opr4"/>
   <feat name="type" value="split"/>
   <feat name="sentenceref" value="p2s3"/>
</struct>
```

b
```
<profileDesc>
   <translations>
      <translation wsd="utf-8" trans.loc="natural-s.xml"/>
      <translation wsd="utf-8" trans.loc="strong-s.xml"/>
   </translations>
</profileDesc>
</cesHeader>
<linkList>
...
   <linkGrp id="p2">
...
      <link>
         <align xlink:href="#p2s3"/>
         <align xlink:href="#xpointer(id('p2s3')/range-to(ids('p2s4')))"/>
      </link>
```

Fig. 4. Output XCES files for the example in Figure 3

Figure 4 shows excerpts of the two new files that were added in this project: (a) the simplification operations and (b) the alignment between natural and strong simplified sentences.

In Figure 4-a, one simplification operation is performed in the sentence identified as p2s3: the operation split. Figure 4-b shows that there is an alignment between p2s3 in natural-s.xml (the XCES file with the natural simplified sentences) and p2s3 and p2s4 in strong-s.xml file (the XCES file with the strong simplified sentences).

In order to align the sentences from the original and simplified versions of the text, we define a cardinality property for each operation, that is, how many sentences should be produced by such operation. The operation of joining sentences has cardinality -1; dropping one sentence has cardinality 0; sentence splitting requires asking the annotator for such cardinality, since different numbers of new sentences may be produced; for all other operations, the cardinality is 1. The cardinality information is used to generate links among original and simplified sentences.

3.3 The Parallel Corpus of Original and Simplified Versions

The first corpus simplified in the PorSimples project is composed of 104 texts from the *Zero Hora* newspaper. These texts were selected because they had a corresponding simplified version, also published in that newspaper, meant to be read by children. Therefore, this parallel corpus can also be useful to evaluate the proposed simplification operations for automatically generating newspaper versions for children. The corpus was simplified by a linguist, expert in text simplification, with the help of the Simplification Annotation Editor, which has been considered user-friendly by the annotator.

Table 2 shows the total number of sentences and words and the average sentence length (in words) of the original, natural and strong simplified texts. The last column shows the percentage of change in the numbers from original texts to strong simplifications. A considerable reduction happened with respect to individual sentence lengths. The overall text length is longer than the original, which was expected, as simplification usually yields the repetition of information in different sentences, particularly when splitting operations are performed. In the PorSimples project, we also provide summarization tools to shorten the texts, as part of the simplification process.

Table 2. Statistics on the original, natural and strong corpora

	Original	Natural	Strong	Change from original to strong
Number of sentences	2,116	3,104	3,537	+ 67.15%
Number of words	41,897	43,013	43,676	+ 4.24%
Average sentence length	19.8	13.85	12.35	- 37.63%

Tables 3 and 4 show the number of sentences, the percentage of sentences with respect to the input texts (original and natural, respectively), and the average sentence length (in words) after the simplifications from *original to natural*, and from *natural to strong*, focusing on two aspects: the types of operations applied and the syntactic phenomena addressed. The total number of sentences in the original corpus was 2,116, with an average sentence length of 19.8 words. The natural simplified corpus resulted in 3,104 sentences, with an average sentence length of 13.86 words. As mentioned before, the number of sentences increases with simplification, but these sentences are usually shorter.

Table 3. Statistics on the simplification operations

Syntactic and Lexical Simplification Operations	Number of sentences / (%) / Average sentence length					
	Original to Natural			Natural to Strong		
Non-simplification	418	19.75%	13.1	2,220	71.52%	11.86
Strong rewriting	7	0.33%	19.85	4	0.13%	14.5
Simple rewriting	509	24.05%	21.91	313	10.0%	16.95
Subject-verb-object ordering	31	1.46%	25.06	13	0.42%	14.15
Transformation to active voice	89	4.21%	22.12	65	2.09%	18.95
Inversion of clause ordering	191	9.03%	22.36	74	2.38%	18.89
Splitting sentences	723	34.17%	26.80	380	12.24%	23.58
Joining sentences	5	0.24%	10.83	6	0.19%	18.33
Dropping one sentence	6	0.28%	11	3	0.09%	5.3
Dropping sentence parts	241	11.39%	26.20	49	1.58%	22.20
Lexical Substitution	980	46.31%	23.46	196	6.34%	18.01

In Table 3, only the “Non-simplification” and “Dropping one sentence” operations are exclusive. The other operations can be combined in one sentence. In the natural simplification process, the most common operation is lexical simplification, followed by splitting sentences, dropping parts of the text, and changing discourse markers by simpler and/or more frequent ones. Strong simplifications (from natural simplifications) prioritize splitting sentences and lexical substitution. The higher number of non-simplification operations in the strong simplification process is due to the fact that most of the sentences had already been simplified in the natural simplification process.

Table 4. Statistics on the syntactic phenomena

Syntactic Phenomena	Number of sentences / (%) /Average sentence length					
	Original to Natural			Natural to Strong		
Apposition	196	9.26%	28.48	54	1.74%	22.20
Coordinate Clauses	806	38.09%	25.31	801	25.80%	18.9
Passive Voice	198	9.35%	26.06	146	4.70%	18.4
Relative Clauses	521	24.62%	25.43	412	13.27%	20.22
Subordinate Clauses	452	21.36%	25.5	524	16.88%	20.03

As shown in Table 4, certain syntactic phenomena are more frequent than others, and therefore many more simplification operations on sentences containing those types of phenomena were performed. The most frequent ones are coordinate, relative and subordinate clauses. These are in general the most difficult cases to simplify, according to studies performed in our project, and we consider this as an additional motivation for the construction of tools to support the simplification process.

4 Conclusions and Future Work

In this paper we have presented a Simplification Annotation Editor and the first corpus resulting from the use of this tool in the context of the PorSimples project. The Editor was developed to help building a parallel corpus of original texts and two simplified versions: natural and strong. Although our focus was on building and analyzing a corpus of newspaper texts, the Editor and the Portal of Parallel Corpora of Simplified Texts can be used to build and query, respectively, other parallel corpora of original and simplified texts from different text genres. For different languages, the language-dependent resources have to be provided and integrated (i) a parser, (ii) a list of simple words, and (iii) dictionaries of complex/ambiguous to simpler discourse markers.

The parallel corpus containing 104 pairs of original and simplified versions can be queried and/or downloaded through the Portal of Parallel Corpora of Simplified Texts to be used in studies of text simplification. Another contribution of this work is the XCES annotation standard for parallel corpora of original-simplified texts, which can also be accessed in the Portal. This corpus can serve as training data for statistical or machine learning methods of simplification; indeed, this work is underway in the PorSimples project.

To summarize, besides the Editor, the PorSimples project has produced the following main contributions: (i) the original-simplified parallel corpora, (ii) the XCES annotation standard developed to register the simplification information and (iii) the Portal of Parallel Corpora to store and query the original or simplified texts.

Our efforts consist of the first step towards the development of automatic text simplification systems for poor literacy readers and potentially people with other cognitive disabilities. The ultimate goal is to help changing the alarming scenario in Brazil, where the majority (68%) of the 30.6 million people between 15 and 64 years who have studied up to 4 years only reach the rudimentary level of literacy, and the majority (75%) of people who studied up to 8 years is only literate at the basic level.

As future work, we will use the resulting corpus to help in the development of rule-based and corpus-based simplifications systems, starting from deciding if a sentence should be simplified or not (non-simplification), and when it should be split, since these cases present a large number of examples.

References

1. Ribeiro, V. M.: Analfabetismo e alfabetismo funcional no Brasil. In: Boletim INAF. Instituto Paulo Montenegro, São Paulo (2006)
2. Max, A.: Writing for Language-impaired Readers. In: Proceedings of Seventh International Conference on Intelligent Text Processing and Computational Linguistics, Mexico City, Mexico. Springer-Verlag, Berlin Heidelberg New York (2006) 567-570
3. Petersen, S. E.: Natural Language Processing Tools for Reading Level Assessment and Text Simplification for Bilingual Education. PhD thesis, University of Washington (2007)
4. Siddharthan, A.: Syntactic Simplification and Text Cohesion. PhD thesis, University of Cambridge (2003)
5. Devlin, S., Unthank, G.: Helping aphasic people process online information. In: Proceedings of the ACM SIGACCESS 2006, Conference on Computers and Accessibility, Portland, Oregon, USA (2006) 225-226
6. Klebanov, B., Knight, K., Marcu, D.: Text Simplification for Information-Seeking Applications. In: On the Move to Meaningful Internet Systems. Volume 3290 of LNCS, Springer-Verlag, Berlin Heidelberg New York (2004) 735-747
7. Vickrey, D., Koller, D.: Sentence Simplification for Semantic Role Labeling. In: Proceedings of the ACL-HLT (2008) 344-352
8. Chandrasekar, R., Srinivas, B.: Automatic Induction of Rules for Text Simplification. Knowledge-Based Systems, 10 (1997) 183-190
9. Daelemans, W., Hothker, A., Sang, E. T. K.: Automatic Sentence Simplification for Subtitling in Dutch and English. In: Proceedings of the 4th International Conference on Language Resources and Evaluation, Lisbon, Portugal (2004) 1045-1048
10. Petersen, S. E., Ostendorf, M.: Text Simplification for Language Learners: A Corpus Analysis. In: Proceedings of the Speech and Language Technology for Education Workshop (SLaTE-2007), Pennsylvania, USA (2007) 69-72
11. Muller, C., Strube, M.: Multi-Level Annotation in MMAX. In: Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue, Sapporo, Japan (2003)
12. Ide, N., Romary, L.: International standard for a linguistic annotation framework. Journal of Natural Language Engineering, 10 (3-4) (2004) 211-225
13. Suderman, K., Ide, N.: Layering and Merging Linguistic Annotations. In: Proceedings of EACL Workshop "Multi-dimensional markup in NLP", Trento, Italy (2006) 89-92
14. Megyesi, B. B., Dahlqvist, B.: The Swedish-Turkish Parallel Corpus and Tools for its Creation. In: Proceedings of NoDaLida 2007, Tartu, Estonia (2007)
15. Specia, L., Aluisio, S. M., Pardo, T. A. S.: Manual de Simplificação Sintática para o Português. Technical Report NILC-TR-08-06. São Carlos-SP (2008) (In Portuguese)
16. Aluísio, S,. Specia, L., Pardo, T., Maziero, E., Caseli, H. M., Fortes, R. "A Corpus Analysis of Simple Account Texts and the Proposal of Simplification Strategies: First Steps towards Text Simplification Systems " In the proceedings of The 26th ACM Symposium on Design of Communication (SIGDOC 2008), pp. 15-22.
17. Williams S., Reiter E.: Generating Readable Texts for Readers with Low Basic Skills. In: Proceedings of ENLG-2005 (2005) 140-147
18. Biderman, M. T. C.: Dicionário Iustrado de Português. Editora Ática, São Paulo (2005)
19. Janczura, G. A., Castilho, G. M., Rocha, N. O.: Normas de concretude para 909 palavras da língua portuguesa. Psic.: Teor. e Pesq. 23 (2007)195-204
20. Pardo, T. A. S., Nunes, M. G. V.: Review and Evaluation of DiZer - An Automatic Discourse Analyzer for Brazilian Portuguese. In: Proceedings of PROPOR 2006. Volume 3960 of LNCS, Springer-Verlag, Berlin Heidelberg New York (2006) 180-189
21. Bick, E.: The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. PhD thesis, Aarhus University (2000)