A Corpus Analysis of Simple Account Texts and the Proposal of Simplification Strategies: First Steps towards Text Simplification Systems

Sandra M. Aluísio[1], Lucia Specia[1], Thiago A. S. Pardo[1], Erick G. Maziero[1], Helena M. Caseli[1], Renata P.M. Fortes[2]
[1]Núcleo Interinstitucional de Lingüística Computacional (NILC), [2]Intermídia Lab
Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo
Av. Trabalhador São-carlense, 400. 13560-970 - São Carlos/SP, Brasil
sandra@icmc.usp.br, lspecia@gmail.com, taspardo@icmc.usp.br, {egmaziero, helenacaseli}@gmail.com, renata@icmc.usp.br

ABSTRACT
In this paper we investigate the main linguistic phenomena that can make texts complex and how they could be simplified. We focus on a corpus analysis of simple account texts available on the web for Brazilian Portuguese (BP). This study illustrates the need for text simplification to facilitate accessibility to information by poor readers and by people with cognitive disabilities. It also highlights features of simplification for BP, which may differ from other languages. Moreover, we propose simplification strategies and a Simplification Annotation Editor. This study consists of the first step towards building BP text simplification systems. One of the scenarios in which these systems could be used is that of reading electronic texts produced, e.g., by the Brazilian government or by news agencies.

Categories and Subject Descriptors
H.3.1 [Content Analysis and Indexing]: *Linguistic processing, Abstracting methods.* H.5.2 [User Interfaces]: *Natural language, Evaluation/methodology.*

General Terms
Design, Human Factors Experimentation

Keywords
Text Simplification, Corpus Analysis, Natural Language Processing.

1.INTRODUCTION
The term *letramento* (literacy) is used in Brazil to designate people's ability to effectively use their reading and writing skills in several aspects of their social life [1]. The INAF index (National Indicator of Functional Literacy) has been annually computed to measure the levels of literacy of the Brazilian population, according to four levels of literacy: (1) **illiteracy**: inability to perform simple tasks involving the decoding of words and phrases; (2) **rudimentary literacy level**: ability to find explicit information in short texts, such as advertisements or short letters; (3) **basic literacy level**: ability to find information in slightly longer texts and also make simple inferences; and (4) **advanced literacy level**: ability to read long texts, find multiple types of information, compare different texts, and perform inferences. According to INAF, a large number of the Brazilian population belong to the rudimentary and basic literacy levels

SIGDOC'08, September 22–24, 2008, Lisbon, Portugal.

Some of the distinguishing features of the literacy levels are the abilities to deal with texts of different lengths and to find information and make associations among them. The Natural Language Processing (NLP) tasks known as Automatic Text Summarization (see, e.g., [2]) and Discourse Parsing (see, e.g., [3], and [4]) handle such topics. The main distinguishing feature is the complexity of the texts itself, which is addressed by the field of Text Simplification (TS). TS aims at providing human readers (and also systems) with a better understanding of a written text through its simplification. In some approaches, this involves both lexical and syntactic structures, by substituting uncommon words by usual words, and by breaking down and changing the syntax of the sentences, respectively [5][6]. Other approaches to TS involve dropping parts from the text and adding extra material to explain difficult terms [7] as well as making it more natural, by addressing the generation of cohesive texts [8].

The project PorSimples (*Simplificação Textual do Português para Inclusão e Acessibilidade Digital*) addresses TS with the goal of building systems to promote the access to Brazilian Portuguese (BP) texts by people at the rudimentary and basic literacy levels, as well as by those with cognitive disabilities (e.g. aphasia and dyslexia). We foresee two systems: (i) an on-line authoring system to help producing simplified texts and (ii) a TS system to allow people to simplify already produced content, particularly on the Web. The TS system will explore not only the areas of summarization, discourse parsing and TS itself, but also text presentation schemes. To the best of our knowledge, there are no TS systems for Portuguese.

In this paper we present a study of the linguistic phenomena that make texts complex. We focus on a corpus analysis of BP simple account texts, that is, texts composed with a certain type of reader in mind, providing, in this case, already naturally simplified texts. We compare these to "normal", i.e., non-simplified, texts. The goal is to illustrate the need for text simplification, highlight simplification characteristics of the Portuguese language, and produce a set of simplification rules, in the form of a manual, for

Portuguese. The results obtained constitute the basis for the implementation of rule-based TS systems and also for the process of corpus annotation to build data-driven approaches to TS. In Section 2 we bring a review of the previous research on TS and present linguistic annotation tools and simplification editors. Section 3 presents our corpus study. The resulting simplification manual for Portuguese grammar is presented in Section 4. Section 5 presents our Simplification Annotation Editor.

2.RELATED WORK

2.1.Text Simplification

It is well known 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 [9][10][11]. The Plain English initiative makes available guidelines (Plain Language) that in principle can be applied for any language. Some recommendations are: write using personal pronouns; keep the subject, verb, and object together; explain only one idea per sentence; use short sentences; use active voice; make syntax simple; use concrete, short, simple words; etc. Although some recommendations are directly useful for TS systems (e.g., subject-verb-object order and active voice), others are difficult to specify (e.g., how simple words and syntactic constructions are). While [9], [11], [12], and [13] only consider lexical and syntactic knowledge to approach TS, [14], [15], [8], and [16] tackle the generation of simplified texts by focusing on choices at the discourse level, trying to answer what choices (e.g., discourse relations, referring expressions, and cue phrases) are most appropriate. [15] and [29] also use psycholinguistic findings on readability as a basis to their easy-to-read text generation system.

The kind of knowledge used to implement TS systems is an important issue and it is related to the use the system is meant for. For example, [12] and [13] design TS methods to produce as output simplifications which are more appropriate to be processed by NLP tools (e.g., a parser is more likely to get a correct structure for a simple sentence than for a complex one), or to be post-processed for human use. [10] focuses on TS to applications on easy information search, producing factoid-like sentences. [17], however, claims that the approach in [10] produces sentences that run the risk of being more difficult to comprehend, as they may have fewer linguistic cues of cohesion that specify how the sentences should be conceptually related. They have developed a tool called Coh-Metrix [34] to measure text complexity. The approach followed by [16], [14], and [7] may be used for educational purposes. Other groups of users may also benefit from TS systems: people using assistive technologies [18][19][20][21], hearing-impaired people who communicate to each other using sign languages like LIBRAS [22]; people with cognitive disabilities caused by medical conditions or interventions [6][23][24][25][26]; and people undertaking distance education [27].

Instead of using TS systems to simplify complex texts, some researchers like [28] defend the use of simple accounts. Simple accounts consist of texts composed in a way that the writer recast the information that he or she abstracts from several sources to suit a particular kind of reader, yielding authentic discourses and more natural texts. While manually generating simple accounts can indeed lead to more natural texts than the automatically simplified ones, this is a very expensive process, which requires specific efforts for different target readers. Moreover, to build deep natural language generation systems for text simplification (see, e.g., [14], [29], and [30]) may be also a complex task, but once a basic framework is defined for automatic TS, variations of these can be relatively easily derived in the form of different systems, tuned to particular readers. In this paper we tackle two subsets of simplification strategies that we call natural and strong simplifications to illustrate how the variations of TS can be addressed. These subsets are described in Section 5 together with the indication of possible users which can benefit from them.

2.2.Support Tools for Linguistic Annotation and Simplification Editors

Linguistic annotation is the process and result of adding new information to existing language data/corpora [37]. Although linguistic annotation is an inherently manual task, some tools have been developed to help humans to perform annotations in a semi-automatic way.

Some tools, such as GATE (http://gate.ac.uk/) and its several systems for English, were developed to automatically annotate a corpus. MMAX (MultiModal Annotation in XML) is other example of a linguistic annotation tool. It is a tool for multi-level annotation of (potentially multi-modal) corpora [37]. However, it is not able to specify relations between different texts, an essential task to the text simplification annotation process, since the relation between an original sentence of a text and its simplified version in another text has to be explicitly specified. There are also some tools called simplification editors, such as SIMPLUS (www.linguatechnologies.com) and StyleWriter (www.editorsoftware.com/writing-software). SIMPLUS is a generic tool for helping writing for simplified (or controlled) English. Simplified English implies the use of limited vocabulary of Standard or Plain English words and restricted sentence structure, but without loss in meaning. StyleWriter has also features to help users to write in Plain English. It guides the user on how to produce a well-written English text and also focus on simplifying and clarifying the text.

The text simplification annotation process in PorSimples project and the proposed Simplification Annotation Editor are presented in Section 5. Some simplification features present in the previous tools are included in our editor. However, instead of helping authors to write simple texts, in the moment our editor is intended to support the creation of a parallel corpus of original-simplified texts to be used in data-driven approaches to TS. Other reasons for creating our own editor are that it must be free and available for the research community and that it is intended to evolve as the project goes on, becoming a TS system itself.

3.A CORPUS ANALYSIS OF SIMPLE ACCOUNTS

It is interesting to notice that simple account texts present texts aligned to visual and meta-linguistic information. They generally use frames, comic strips, balloons, attention-calling phrases, parody, numbered and spaced paragraphs, definitions for difficult words, highlighting for important pieces of information (bold, italic and in bigger sizes), etc.

We conducted a corpus analysis to verify how simple such texts actually are and which characteristics cause them to be natural and authentic. In particular, we wanted to measure how texts could be quantified in terms of the previous work findings on how simple

texts must be. We focused our analysis on the following points: size of the sentences; size of the words; number of relative clauses and appositions; subordinate and coordinate conjunctions and their positions; main and subordinate clause ordering; number of reduced and finite clauses; number of simple words. We analyzed 6 corpora of simple account texts in BP. They belong to different genres and are available on the Web:

(1) Corpus *Ao Encontro da Lei* (hereafter *Enc*), a simple version of Brazilian New Civil Code;

(2) Corpus *Cartilha de Orientação Legal – Brasileiras e Brasileiros no Exterior* (hereafter *Ca*), an effort of the Brazilian government to make available information about living abroad;

(3) Corpus *Bulário da ANVISA* (hereafter *Bu*), composed of easy-to-read medicine directions;

(4) Corpus *De palavra em palavra* (hereafter *Dp*), an initiative from a news agency (*O Estado de São Paulo*) to build texts about Portuguese Grammar accessible to youngsters;

(5) Corpus *Para seu Filho Ler* (from *Zero Hora*) (hereafter *ZH*), which comprises versions of news texts for children;

(6) Corpus *Ciência Hoje das Crianças* (hereafter *CHC*), a magazine initiative to build scientific texts for children.

A non-simple account corpus, *Caderno Brasil da Folha de São Paulo* (hereafter *FSP*), was analyzed, so that its features could be contrasted to those of the simple accounts. It is composed of news about Brazil aimed for a wide audience, collected from corpus PLN-Br GOLD [35], publicly available on the Web. This was chosen to allow the comparison between "normal" and simple account texts. We analyzed 55 simple account texts: 10 sections of corpora (1) and (2), 5 sections of corpus (3), and 10 texts of corpora (4)-(6). For the *FSP* corpus, we selected 12 news articles, following a sampling technique used in PLN-Br GOLD, which contains news from 1994 to 2005. Initially, each corpus was automatically annotated by PALAVRAS, a syntactic parser for Portuguese [36]; the corpus analysis was performed by the AIcorpus tool (this is a tool to analyse several features of a corpus, available at www.nilc.icmc.usp.br/AIC).

In order to count the simple words in each corpus, we used a previously compiled list. The discourse markers counted were those identified by [4] for BP. Table 1 lists the total number of sentences and words, average sentence length and the percentage of 'simple' words in each of the seven corpora. One may see that all the 6 corpora of simple account texts have fewer words per sentence than the *FSP* corpus, i.e., the non-simple account text. They also contain more common words.

Regarding the size of the words, *FSP* has on average 5.06 characters per word, while *Ca* has 5.61, and the remaining texts have also a similar number of characters per word, on average: from 4.67 to 4.91, that is, close to *FSP*.

Table 1. Simple statistics from the 7 corpora

	# words	Average words per sentence	# sentences	% simple words
ZH	1116	16.91	66	87.9
CHC	4417	19.72	224	88.9
Ca	2633	20.09	131	81.28
Bu	8141	15.86	513	81.19
Dp	2052	15.91	129	81.82
Enc	2161	20.39	106	86.86
FSP	5574	21.11	264	80.97

Table 2 shows the figures resulting from the analysis of several other features in the 7 corpora. Although all the simple account corpora have fewer prepositional phrases and embedded apposition than the *FSP* corpus, contrary to what we expected, we cannot conclude that simple account texts contain less or more relative clauses, passive voice sentences, enumerative apposition, adjectives or adverbs, which are all supposed to make the text more complex. One fact, although, is important to notice: the *Bu* corpus presents a large number of enumerative appositions. We have checked those instances and verified that this construct has strong correlation to the use of a paralinguistic feature – listings with bullets or numbers for several events related to the medicines, e.g., symptoms. As for relative clauses, all the simple account corpora except *Bu* have a large number of them. In *CHC*, they are related to the definition of concepts or terms. Splitting the relative clauses and other complex constructions in two sentences would improve the readability of these texts. This operation is discussed in Section 4. Table 3 shows that the simple account corpora, except those aiming for children, contain proportionally more sentences with only one or two clauses (1 and 2-clause sentences) than *FSP* corpus, i.e., *FSP* seems indeed to contain more complex syntactic constructions.

This finding regarding to the simple accounts aimed to children was curious, as *ZH*, for example, has the smallest number of coordinate and subordinate conjunctions, the smallest number of non-finite verbs and is among the ones with the smallest number of words per sentence, on average. It seems that the most used syntactic construct is the asyndetically coordinate clause, maybe due to the decision to shorten the length of the sentences. Following the recommendation of Plain Language, in all the 7 corpora there is still room for improving sentences readability by splitting the sentences with 3 or more clauses. In particular, readability of the simple account corpora would be improved if the number of initial subordinate clauses were reduced. Analyzing discourse markers, we noticed that there are a larger number of exemplification markers in the simple account corpora than in the *FSP* corpus. The short markers (e.g., *também* (also), *se* (if), *quando* (when), *ou* (or), *como* (as/like), and *bem* (well)) also appear in larger number than in *FSP*. Simple accounts also have a larger number of discourse markers than *FSP* in general (following the order of the corpora in the tables, the percentages are: 10.3, 12.49, 9.08, 10.37, 10.92, 12.22 and 7.30).

Table 2. Prepositional phrases, adjectives, adverbs, relative clauses, apposition, and passive voice in the 7 corpora

	Prepositional Phrases*	Average PPs per sentence	Average PPs per clause	Relational Clauses (%)	Apposition		Passive sentences (%)	Adjectives (%)	Adverbs (%)
					*Embedded**	*Enumerative*			
ZH	16	0.24	0.08	11	1	3	4.55	3.85	13.53
CHC	66	0.29	0.09	18.02	15	19	14.73	6.02	15.24
Ca	37	0.28	0.13	14.95	8	22	6.87	8.81	10.75
Bu	68	0.13	0.09	9.1	5	39	6.43	9.56	12.78
Dp	36	0.28	0.14	13.53	10	23	10.85	5.51	14.18
Enc	42	0.39	0.15	13.07	5	3	15.09	5.55	13.74
FSP	107	0.41	0.15	9.27	22	13	9.85	5.45	11.39

* Incidence calculated in the first 60 sentences of the corpora

Table 3. Clauses in the 7 corpora

	Initial subordinate clause (%)	Initial coordinate clause (%)	1-clause + elliptical clause sentences (%)	2-clause sentences (%)	3-clause sentences (%)	4-clause sentences (%)	5-clause sentences (%)	More than 5-clause sentences (%)	Average clauses per sentence	Coordinating conjunctions (%)	Subordinating conjunctions (%)	Non-finite verbs (%)
ZH	0	3.03	19.67	24.24	22.73	13.64	13.64	7.57	3.03	4.03	2.78	7.52
CHC	5.8	7.59	22.76	21.88	23.21	13.84	6.69	15.17	3.02	3.39	2.47	6.72
Ca	3.05	2.29	36.64	29.77	16.03	7.63	6.87	3.81	2.15	4.86	1.48	5.28
Bu	7.21	0.38	42.88	29.62	13.25	10.13	2.72	1.36	1.94	3.58	1.76	6.84
Dp	5.43	9.3	42.5	24.03	17.83	6.98	4.65	4.66	2.06	3.95	1.80	4.09
Enc	8.49	9.43	32.8	30.19	10.3	13.21	7.54	5.65	2.67	4.16	2.45	7.31
FSP	2.27	21.96	29.54	20.45	25.37	11.74	7.95	4.91	2.66	2.33	2.24	5.16

4.A MANUAL FOR PORTUGUESE SYNTACTIC SIMPLIFICATION

As a result of the studies presented in Section 3, we defined a set of operations related to certain linguistic phenomena, which can be performed on Portuguese texts in order to simplify such texts. This set was compiled in the form of a manual to be used both for the creation of rules in a rule-based text simplification system, and to guide human annotators to simplify texts in order to produce examples to train machine learning techniques to learn such rules. For human use, specifically, such rules were the basis for the development of an annotation tool, which is described in the next section.

As shown in Tables 4 to 9, the manual is organized in 6 sections describing how the syntactic constructs and discourse markers – a lexical choice based on discourse information – should be simplified. In the manual we provide several examples of simplification operations. The six constructs are: apposition, relative clauses, subordinate clauses, passive voice, sentences with non-finite verbs, and coordinate clauses. There are 5 simplification operations: **a)** splitting sentences, **b)** changing a discourse marker by a simpler and/or more frequent one (the indication is to avoid the ambiguous ones), **c)** changing passive to active voice, **d)** inverting clause order and **e)** non-simplification. The general guidelines are: to shorten sentences; keep the subject-verb-object together; to avoid embedded sentence between parentheses, commas, or dashes.

Tables 4 to 9 show, for each kind of construct, the simplification operations to be applied, the suggested order of the clauses, and the cue phrase(s) (translated into English), when it applies. The "comments" column illustrates the general case of the simplification, although there are rules for specific cases of each construct. For an example of simplification operation, consider the following original text: "*The building hosting the Brazilian Consulate was also evacuated, although the diplomats have obtained permission to carry on working*". Its simplified version, applying the rule for concessive subordinate clauses, would be: "*The diplomats have obtained permission to carry on working. But the building hosting the Brazilian Consulate was also evacuated"*. The sentence is split in two, the clauses are inverted, and a simple discourse marker ("But") is chosen.

Table 4. Simplification operations for "Apposition"

Construct	*Op.*	*Order of clauses*	*Cue phrase*	*Comments*
Enumerative	**e**			Used to list items in simple accounts
Embedded (app.)	**a**	Original/ App.		Appositive: Subject is the head of original + to be in present tense + apposition

Table 5. Simplification operations for "Relative Clauses"

Construct	*Op.*	*Order of clauses*	*Cue phrase*	*Comments*
Non-restrictive	**a**	Original/ Relative		Relative: Subject is the head of original + relative
Restrictive	**a**	Relative/ Original		Relative: Subject is the head of original + relative

Table 6. Simplification operations for "Subordinate Clauses"

Construct	*Op.*	*Order of clauses*	*Cue phrase*	*Comments*
Causal /Reason	**a, b, d**	Sub/Main	*With this*	To keep the ordering cause, result
Comparative	**a, b**	Main/Sub	*Also*	Rule for *such ... as*, *so ... as* markers
	e			Rule for the others markers or short sentences
Concessive	**a, b, d**	Sub/Main	*But*	Clause 1 *although* clause 2 is changed to clause 2. *But* clause 1
	a, b	Main/Sub	*This happens even if*	Rule for hypothetical sentences
Conditional	**e**			Pervasive use in simple accounts
Result	**a, b**	Main/Sub	*Thus*	May need some changes in verb
Final /Purpose	**a, b**	Main/Sub	*The goal is*	
Proportional	**e**			Sub. clause frequently appears without a verb
Confirmative	**a, b, d**	Sub/Main	*Confirms that*	May need some changes in verb
Temporal	**a**	Sub/Main		May need some changes in verb
	a, b		*Then*	Rule for the markers: after that, as soon as

Table 7. Simplification operation for "Sentences with non-finite verbs"

Construct	*Op.*	*Order of clauses*	*Cue phrase*	*Comments*
Non-finite verbs	**e**			Used to shorten sentences

Table 8. Simplification operation for "Passive Voice"

Construct	*Op.*	*Order of clauses*	*Cue phrase*	*Comments*
Passive voice	**c**			

Table 9. Simplification operations for "Coordinate Clauses"

Construct	*Op.*	*Order of clauses*	*Cue phrase*	*Comments*
Asyndetic	**a**	Keep order		New sentences: Subjects are the head of the original subject
Additive	**a**	Keep order	*Keep marker*	Keep marker; it appears in the beginning of the new sentence
Adversative	**a, b**	Keep order	*But*	
Correlated	**a, b**	Keep order	*Also*	Original markers disappear
Result	**a, b**	Keep order	*As a result*	
Reason	**a, b**	Keep order	*This happens because*	May need some changes in verb

5.A SUPPORT TOOL TO HELP MANUAL SIMPLIFICATION ANNOTATION

We could realize that readers with different literacy levels need different types of help. A large number of researchers relate the capabilities and performance of the working memory with reading levels (e.g., see [8]). Several studies have also shown that splitting complex sentences results in the reduction of information in the working memory, although such operation may cause the text to look a bit non-cohesive (e.g., see simplifications proposed in [10]). In PorSimples we also want to help poor readers to improve their reading skills over the time. [26], for example, states that understanding and learning through texts are not enhanced when based only on simplified and coherent texts. Although simplification is basically an educational action that all teachers perform every day, this action must be well balanced.

In order to achieve balance we propose two subsets of simplification operations called here natural and strong simplifications. They were designed by observing and analyzing an expert in text revision to simplify a newspaper article in Portuguese: from all the operations related in Tables 4-9 plus lexical simplification and dropping text parts (not covered by the simplification manual), sentence splitting was the only one used with parsimony. The natural simplification subset includes all but splitting operation, while strong simplification involves it. Below we show the first sentence of an article of the FSP newspaper (section Brazil, 2005, translated into English) to illustrate the difference of a natural simplified text from a strongly simplified version:

> In a press conference called to answer corruption charges during his term as Mayor of the city of Ribeirão Preto, the Minister Antonio Palocci Filho (Treasury) said to be willing to resign his position, but with the recommendation of President Luiz Inácio Lula da Silva, would remain in government.

The natural simplified version is shown below (notice that we still have 3 clauses in the first sentence):

> The Minister Antonio Palocci (Treasury) said in a press conference that will leave his position, although President Lula advised him to remain in the government.

while the final strong simplified one (using splitting operation) is:

> The Minister Antonio Palocci is the Treasury Minister. Antonio Palocci said in a press conference that will leave his position. But he said that President Lula advised him to remain in the government.

Based on these ideas, the Simplification Annotation Editor developed at PorSimples follows a three steps architecture, which includes the revision of the source text (original version) when necessary and the two subsets of simplification (natural and strong). In the first step, the human annotator manually revises the source text correcting punctuation and misspellings. Then, the revised source text is used to start the simplification process. The strong simplification step is based on the outcome of the natural simplification step.

The difference between natural and strong simplifications is that, in the first, the human annotator is free to perform simplification without following any specific rule, while in the strong simplification, he/she has to follow pre-defined syntactic simplification rules described in [38].

Figure 1 shows a screenshot of the strong simplification process. The natural simplification screen is very similar. The editor is split in three areas: (1) the text being simplified, (2) the simplified version being produced, and (3) the log of simplification operations performed so far (in this example, for the third original sentence). The simplification operations are accessible by a pop-up menu as shown in the figure. They encompass all the operations from Tables 4-9 except the operation b (such operation is performed in the *Léxico* mode, which is defined latter).

Thus, the 10 simplification operations in the menu are: (1/e) non-simplification; (2) simple or (3) strong rewriting (as defined in [7]); (4) putting the sentence in its canonical order (subject-verb-object); (5/c) putting the sentence in the active voice; (6/d) inverting the clause ordering; (7/a) splitting or (8) joining sentences; (9) dropping one sentence and (10) dropping sentence parts. The editor has two auxiliary modes to help the human annotator to decide when to perform these operations: the *Léxico* and the *Sintático* modes. In the *Léxico* mode, the editor proposes changes in words and discourse markers (operation b from Tables 4-9) by simpler (or non-ambiguous) and/or more frequent ones. These lexical simplification operations 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 [32], frequent words from news texts for children, and concrete words [33]. The discourse markers were defined based in [4].

The *Sintático* mode, in turn, proposes syntactic operations such as splitting sentences in those cases summarized in Tables 4-9 and detailed in [38]. The syntactic operations are proposed based on the syntactic information about the text provided by a parser for Portuguese [36].

As an example, in Figure 1, the system recommends (in the small recommendation box) splitting the third original sentence, since it has a coordinate clause (see Table 9 and the coordinate conjunction "*e*" in different background color). This operation can be selected in the recommendation box or in the menu, and is shown in the area 3 that exhibits the log of simplification operations. For each simplification operation (in the area 3) it is possible to specify (by means of "*Detalhar operação*") what has been changed in the simplified version. In the example given (in Brazilian Portuguese), the natural simplified third sentence (in the area 1):

No filme, um cardume de piranhas escapava de um laboratório militar e atacava participantes de um festival aquático.

was split into two sentences as shown below:

No filme, um cardume de piranhas escapava de um laboratório militar. O cardume de piranhas atacava participantes de um festival aquático.

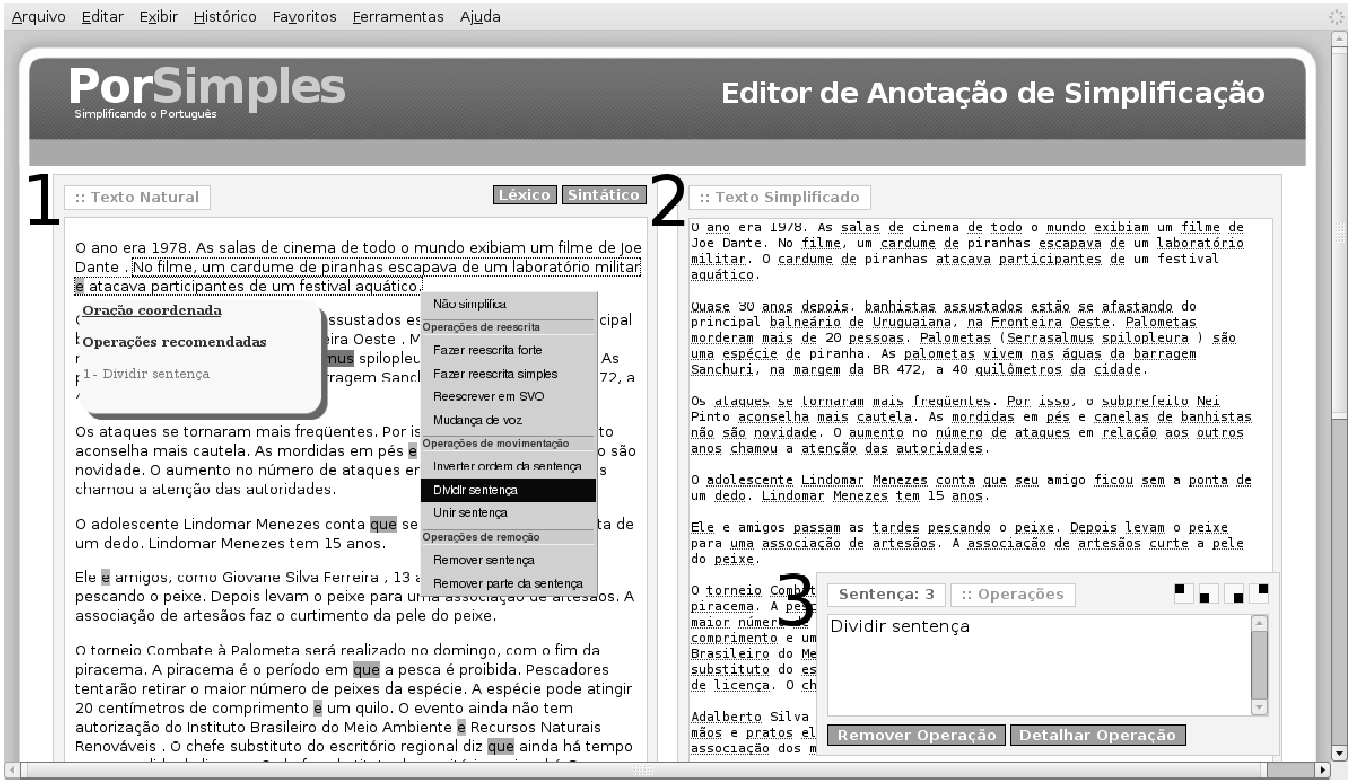

Figure 1. Screenshot of the Simplification Annotation Editor (in the *Sintático* mode)

6.FINAL REMARKS AND FUTURE WORK

We presented in this paper the first steps towards producing TS systems for BP texts under the PorSimples project, which aims at allowing poor readers to have easier access to information. From the study, we could verify that TS is a necessary task and that even simple account texts could be more tuned to their final usage. The study also gave rise to the first syntactic simplification manual for BP and the grouping of the simplification operations in two subsets: natural and strong simplifications. The manual and

the Simplification Annotation Editor will serve as a basis for annotating corpora and for producing automatic TS systems, which consist in the immediate future work we foresee in the project. Text summarization and information elicitation tasks are also under investigation for TS purposes.

7.ACKNOWLEDGMENTS

We thank FAPESP and Microsoft Research for supporting the PorSimples project. The authors would like to thank Carol Scarton, Tiago Pereira, Rachel Aires, Amanda Rocha, Arnaldo Candido Jr. and Paulo Margarido for their valuable work on corpus compilation, and analysis and validation of the operations of the syntactic simplification manual for BP; also Tiago Pereira and Felipe Perez for the design, implementation and evaluation of the Simplification Annotation Editor.

8.REFERENCES

[1] Ribeiro, V. M.: Analfabetismo e alfabetismo funcional no Brasil. Boletim INAF. São Paulo: Instituto Paulo Montenegro (2006)

[2] Rino, L.H.M., Pardo, T.A.S., Silla Jr., C.N., Kaestner, C.A., Pombo, M.: A Comparison of Automatic Summarization Systems for Brazilian Portuguese Texts. SBIA 2004, LNAI, vol. 3171, pp. 235-244. Springer, Heidelberg (2004)

[3] Feltrim, V., Pelizzoni, J.M., Teufel, S., Nunes, M.G.V., Aluísio, S.M.: Applying Argumentative Zoning in an Automatic Critiquer of Academic Writing. SBIA 2004, LNAI, vol. 3171, pp. 1-10. Springer, Heidelberg (2004)

[4] Pardo, T.A.S., Nunes, M.G.V.: Review and Evaluation of DiZer - An Automatic Discourse Analyzer for Brazilian Portuguese. PROPOR 2006, LNCS, vol. 3960, pp. 180-189. (2006)

[5] Mapleson, D.L.: Post-Grammatical Processing for Discourse Segmentation. PhD Thesis. School of Computing Sciences, University of East Anglia, Norwich (2006)

[6] Max, A.: Writing for Language-impaired Readers. In the Proceedings of Seventh International Conference on Intelligent Text Processing and Computational Linguistics. CICLing 2006, pp. 567-570. (2006).

[7] Petersen, S. E., Ostendorf, M.: Text Simplification for Language Learners: A Corpus Analysis. Speech and Language Technology for Education workshop, October 2007, Pennsylvania, USA. Available at: www.sarahpetersen.net/portfolio/Petersen_Ostendorf_SLaTE2007_final.pdf (2007)

[8] Siddharthan, A. Syntactic Simplification and Text Cohesion. PhD Thesis. University of Cambridge (2003)

[9] Siddharthan, A.: An Architecture for a Text Simplification System. In the Proceedings of the Language Engineering Conference (LEC), pp. 64-71. (2002)

[10] Klebanov, B., Knight, K., Marcu, D.: Text Simplification for Information-Seeking Applications. On the Move to Meaningful Internet Systems. LNCS, vol. 3290, pp. 735-747. Springer-Verlag (2004)

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

[12] Chandrasekar R., Doran C. and Srinivas, B.: Motivations and Methods for Text Simplification. COLING 1996, pp. 1041-1044. (1996)

[13] Chandrasekar, R., Srinivas, B.: Automatic induction of rules for text simplification. Knowledge-Based Systems, 10, 183–190. (1997)

[14] Williams, S.: Natural Language Generation (NLG) of discourse relations for different reading levels. PhD Thesis, University of Aberdeen. (2004)

[15] Williams, S., Reiter, E.: A corpus analysis of discourse relations for Natural Language Generation Proceedings of Corpus Linguistics 2003, Lancaster University pp. 899-908. (2003)

[16] Siddharthan, A.: Syntactic Simplification and Text Cohesion. Research on Language and Computation 4:77–109. Volume 4, Number 1 / June, (2006)

[17] McNamara, D.S., Louwerse, M.M., Graesser, A.C.: Coh-Metrix: Automated cohesion and coherence scores to predict text readability and facilitate comprehension. Grant proposal. Available at: http://cohmetrix.memphis.edu/cohmetrixpr/publications.html (2002)

[18] Cook, A.M. , Hussey, S.M.: Assistive Technologies: Principles and Practice. Mosby (1995)

[19] Freire, A.P., Paiva, D.M.B., Turine, M.A.S., Fortes, R.P.M.: Using Screen Readers to Reinforce Web Accessibility Education. In the Proceedings of the 12th ACM Annual Symposium on Innovation and Technology in Computer Science Education. pp. 82-86. ACM Press. (2007)

[20] Freire, A.P., Fortes, R.P.M.: Automatic accessibility evaluation of dynamic web pages generated through XSLT. In the Proceedings of the 2005 International Cross-Disciplinary Workshop on Web Accessibility, pp. 81-84. ACM Press. (2005)

[21] Freire, A P., Goularte, R., Fortes, R. P. M.: Techniques for Developing More Accessible Web Applications: a Survey Towards a Process Classification. In: The Proceedings of 25th ACM International Conference on Design of Communication, pp. 162-169. ACM Press. (2007)

[22] Meireles, V., Spinillo, A.G.: Uma análise da coesão textual e da estrutura narrativa em textos escritos por adolescentes surdos. Estudos de Psicologia, V. 9, N. 1, pp. 131-144. (2004)

[23] Inui, K.; Fujita, A., Takahashi, T., Iida, R., Iwakura, T.: Text simplification for reading assistance: a project note. In the Proceedings of the Second International Workshop on Paraphrasing, pp. 9-16. Sapporo, Japan. (2003)

[24] Daelemans, W., Hothker, A., Sang, E.T.K.: Automatic Sentence Simplification for Subtitling in Dutch and English., LREC 2004, pp. 1045-1048. (2004)

[25] Carroll, J., Minnen, G., Canning, Y., Devlin, S., Tait, J.: Practical simplification of English newspaper text to assist aphasic readers. In the Proceedings of AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology. (1998)

[26] Gordon, W.: The Interface Between Cognitive Impairments and Access to Information Technology. In S. Keates (ed), Accessibility and Computing. ACM Special Interest Group on Accessible Computing, V. 83, pp. 3-6. (2005)

[27] Ramos, W. M.: A compreensão leitora e a ação docente na produção do texto para o ensino a distância. Linguagem & Ensino, Vol. 9, No. 1, pp. 215-242. Universidade de Brasília. (2006)

[28] Widdowson, H. G.: Teaching language as communication. Oxford: Oxford University Press. (1978)

[29] Williams S., Reiter E.: Generating basic skills reports for low-skilled readers. To appear in Natural Language Engineering. In press. (2008)

[30] Williams S., Reiter E.: Generating Readable Texts for Readers with Low Basic Skills. Proceedings of ENLG-2005, pp. 140-147. (2005)

[31] Carvalho Netto, J. R.: Ao Encontro da Lei: O Novo Código Civil ao alcance de todos. São Paulo: Imprensa Oficial. (2003)

[32] Biderman, M. T. C. DICIONÁRIO ILUSTRADO DE PORTUGUÊS. São Paulo, Editora Ática. 1ª. ed. São Paulo: Ática. (2005)

[33] Janczura, G. A., Castilho, G. M., Rocha, N. O.: Normas de concretude para 909 palavras da língua portuguesa. Psic.: Teor. e Pesq. [online]., vol. 23, pp. 195-204. (2007)

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

[35] Muniz, M., Paulovich, F. V., Minghim, R., Infante, K., Muniz, F., Vieira, R., Aluísio, S.: Taming the tiger topic: an XCES compliant corpus Portal to generate subcorpus based on automatic text topic identification. In: Proceedings of the Corpus Linguistics Conference. pp. 1-18 (2007)

[36] Bick, E.: The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. PhD thesis. Aarhus University. Denmark University Press. (2000)

[37] Muller, C., Strube, M.: Multi-Level Annotation in MMAX. In Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue, Sapporo, Japan. (2003)

[38] 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)