Towards Brazilian Portuguese Automatic Text Simplification Systems

Sandra M. Aluísio[1], Lucia Specia[1], Thiago A. S. Pardo[1], Erick G. Maziero[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@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 and propose simplification strategies for this language. This study illustrates the need for text simplification to facilitate accessibility to information by poor literacy readers and potentially by people with other cognitive disabilities. It also highlights characteristics of simplification for Portuguese, which may differ from other languages. Such study consists of the first step towards building Brazilian Portuguese 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 relevant 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, Poor Literacy Readers, Brazilian Portuguese.

1. INTRODUCTION

In Brazil, letramento (literacy) is a term used to designate people's ability to use written language to obtain and record information, express themselves, plan and learn continuously, i.e., to effectively use their reading and writing skills in several aspects of their social life [1]. Since 2001, the INAF index (National Indicator of Functional Literacy) has been annually computed to measure the levels of functional illiteracy of Brazilian population. INAF 5-year report identifies four levels of literacy for Brazilian population:

(1) **Illiteracy**: the condition of those who are not able to perform simple tasks involving the decoding of words and phrases;

(2) **Literacy – rudimentary level**: the ability to find explicit information in short texts, such as advertisements or short letters;

(3) **Literacy – basic level**: the ability to find information in slightly longer texts and also make simple inferences;

(4) **Literacy – advanced level**: the ability to read long texts, find multiple types of information, compare different texts, and perform inference.

The average scores obtained in the exams applied show that the proportion of adult people with higher education levels has been increasing (people with high school or higher levels increased from 28% in 2001 to 36% in 2005). However the average performance in each education level shows a negative drift. In fact, according to INAF, the majority (68%) of the 30.6 million Brazilians between 15 and 64 years who have studied up to 4 years only reach the rudimentary level of literacy. Amongst the people who studied for 8 years, only a quarter can be considered fully literate, while the vast majority is literate at the basic level.

One of the relevant features in the three levels of literacy is the ability to deal with texts of different lengths. This feature can be addressed by a very well known Natural Language Processing (NLP) task – automatic summarization – which can be applied to original texts in order to generate new texts with different degrees of compression (see, e.g., [2]). Another feature is the ability to find information (e.g., text purpose, context and conclusions) and make associations among parts of the text (e.g., contrasts, exemplifications and cause-effect associations). This ability is usually addressed by the field of automatic discourse analysis (e.g., [3, 4]). The main distinguishing feature in the three levels of literacy refers to the complexity of the texts itself, which is addressed by the field of Text Simplification (TS).

TS is an application of an emerging area of research in the field of Natural Language Processing (NLP) called text-to-text generation. TS aims to maximize the comprehension of written texts through the simplification of their linguistic structure. These simplifications may

DocEng'08, September 16–19, 2008, São Paulo, Brazil.

involve lexical and syntactic structures by substituting words that are only understood by a small number of people by words that are more usual, and by breaking down and changing the syntax of the sentences, respectively. As a result, it is expected that the whole text can be more easily understood both by human readers and computer systems [5, 6]. Other approaches to TS may also involve dropping parts from the text and adding extra material to explain difficult terms [7] as well as to make it flow more naturally by addressing the generation of cohesive texts. [8] is an example of the latter approach: it considers sentence ordering, cue-word selection, referring-expression generation, determiner choice and pronominal use during syntactic simplification.

The project PorSimples (*Simplificação Textual do Português para Inclusão e Acessibilidade Digital*) addresses TS aiming at building systems to promote the access to Brazilian Portuguese texts by people at the rudimentary and basic literacy level, 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 read Web content (working as a post-processing system). The latter will explore not only the tasks of summarization, discourse analysis 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. Since there is no corpus of original- simplified text pairs readily available for Portuguese, we focus on a corpus analysis of Brazilian Portuguese simple account texts available on the Web to compare them to "normal", i.e., non-simplified, texts and learn how to make more natural simplifications from them. Simple accounts consist of texts composed in a way that the writer recasts the information that he or she abstracts from several sources to suit a particular kind of reader, yielding authentic discourses and more natural 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 short review of the previous research on TS. Section 3 presents several sources used to help the design of our simplification operations, including the corpus study. The resulting simplification manual for six constructs of the Portuguese grammar is presented in Section 4.

2. RELATED WORK

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]. Some of these problems have been addressed in different ways in the previous work on TS. [12, 13] consider only syntactic knowledge to approach TS, using both rule-based systems and rules learned from a corpus, respectively. [14, 15, 8, 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 for people with poor literacy. The PSET (Practical Simplification of English Texts) project [11] investigated how lexical-level and syntactic level choices affect readability for a special kind of readers – aphasics – without considering discourse choices.

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, 13] design TS methods to produce as output simplifications which are more appropriate to be processed by language processing 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] focus on TS to applications to facilitate information search. They define the concept of Easy Access Sentence (EAS) using the following requirements: (i) EAS is a grammatical sentence; (ii) it has one finite verb; (iii) it does not make claims that were not present in the original text; and (iv) the more named entities a sentence satisfying the previous 3 requirements has, the better EAS it is. An example of text simplification using EAS given in [10] is the following:

> Harriet Beecher Stowe is a writer. Harriet Beecher Stowe was born in Litchfield, Connecticut, USA. Harriet Beecher Stowe is the daughter of Lyman Beecher. Harriet Beecher Stowe was raised by her severe Calvinist father. Harriet Beecher Stowe was raised by Lyman Beecher. Lyman Beecher is Harriet Beecher Stowe's father. Harriet Beecher Stowe was educated at the Hartford Female Seminary. Harriet Beecher Stowe taught at the Hartford Female Seminary. Catherine Beecher founded the Hartford Female Seminary. Catherine Beecher is Harriet Beecher Stowe's sister.

This simplification was produced from a one paragraph biography of Harriet Beecher Stowe, as follows:

> Harriet Beecher Stowe is a writer. She was born in Litchfield, Connecticut, USA, the daughter of Lyman Beecher. Raised by her severe Calvinist father, she was educated and then taught at the Hartford Female Seminary (founded by her sister Catherine Beecher).

The factoids this simplification method generates (subject-verb-objects and possibly some modifiers) makes it easier to retrieve the information in the text. However, it makes the text longer and flat, probably less interesting for human readers. Moreover, [17] claim that EAS-like sentences 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. The approach followed by [16, 14, 7] favors text accessibility to a wider audience of readers, and may be used for educational purposes.

Besides poor literacy readers, which are the focus of our project, other groups of users may benefit from TS systems:

- people making use of assistive technologies, such as screen readers and translators [18, 19, 20, 21];
- hearing-impaired people who communicate to each other using sign languages like LIBRAS (Brazilian Sign Language), since the structural differences between LIBRAS and Portuguese make it difficult to understand complex texts [22];
- people with cognitive disabilities caused by medical conditions or interventions, e.g., people suffering from aphasia or dyslexia [23, 24, 6, 25] and traumatic brain injuries, strokes and aneurysms [26]; and
- people undertaking Distance Education, in which text understandability is of great importance [27].

Instead of using TS systems to simplify complex texts, some researchers like [28] defend the use of simple accounts. One example of simple accounts in Portuguese is the book of legal guidance called *Ao Encontro da Lei: O Novo Código Civil ao Alcance de Todos* [31], which is a simple version of some chapters of the New Brazilian Civil Code. Below we show a text span from the New Brazilian Civil Code:

> Article 16. Every person has the right to a name, in which is included a first name and surname. ... Article 19. The alias chosen for legal activities has the same protection as given to the name.

and the corresponding content from the simple account book (the graphical presentation of the text, i.e., the use of short lines that go up to half of the page for easier viewing , was preserved):

> *João Brasil and Maria Brasil are choosing*
> *the name of their next kid, who is about to be born.*
> *Every person has the right to have a name. And the*
> *law protects people's names. Also, the*
> *law protects a person's alias, like*
> *what happens to Xuxa, Pelé.*
> *The name is made up of a first name and a surname*
> *(name = first name + surname). For*
> *example, João is the first name, Brasil is the surname.*
> *Sometimes, the first name is compound, like*
> *Antônio Carlos, Maria do Carmo. And,*
> *also, the surname can be compound.*
> *For example: Carvalho da Silva.*

Ao Encontro da Lei exploits stories in famous comics, parodies (facts from Brazilian soap operas or movies), subtitles, many definitions, usually through explanations, and reformulations near to difficult words. Sentences are short, but not always composed of a single clause as in EAS. Discourse markers are pervasive, appearing usually at the beginning of the sentences. 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 dedicated efforts for different target readers. On the other hand, while building deep natural language generation systems for text simplification (see, e.g., [14, 29, 30]) is also a complex task, 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 strategies can be addressed. These subsets are described in Section 4 together with the indication of possible users which can benefit from them.

3. DESIGNING SIMPLIFICATION OPERATIONS

Two characteristics of texts that are of interest in this paper are the legibility (graphical presentation of the text) and readability (use of frequent words and simple syntactic structures). Besides the microstructure, the macrostructure of the text is also a concern, in which other features can act as facilitators to understand the text, e.g. the organization, cohesion, coherence, and the focus on particular target readers. For example, the author can bring an anaphora close to its referent, use discourse markers between sentences, give preference for explicit definitions or use complete information. The next three subsections present the sources of information used in the design of a manual for Brazilian Portuguese syntactic simplification.

3.1 Plain Language

Plain English is a movement in Britain and the USA that emerged in the late 1970's as a reaction to the unclear language used in government and business forms and documents. It provides guidelines (the *Plain Language*) that in principle can be applied for any language. Some recommendations on how to write and organize information in Plain Language are: write using personal pronouns; use a simple logic to create connections between obvious ideas; remove all the information that is not essential for the purpose of the text; use a summary for large documents or create a short introduction to the content of each item; keep the subject, verb and object together; explain only one idea per sentence; use short sentences; avoid hidden verbs; use active voice; make syntax simple; use no more than two or three subordinate levels in one sentence; if possible, use the word "if" for conditions; use concrete, short, simple words; avoid or explain legal, foreign, and technical jargon; minimize abbreviations; place the main idea before exceptions and conditions.

Although some recommendations are directly useful and can be implemented in TS systems (e.g. subject-verb-object order, active voice and subordinate clauses control), others are difficult to specify (e.g., how simple each syntactic construction is and which words are simple). Therefore, explicit syntactic simplification rules and lists of simple words are necessary. For English, some lexical resources are available, like the MRC Psycholinguistic Database (which helps to identify difficult words using psycholinguistic measures), but such resources do not exist for Portuguese. For the PorSimples project, we have compiled a list of simple words composed of words supposed to be common to youngsters (from [32]), a list of frequent words (from news texts for children) and a list of concrete words available in [33]. We have also defined a set of syntactic simplification guidelines, as we describe in Section 4.

3.2 TS Systems for English and Coh-Metrix

Siddharthan [9] illustrates the simplification of various syntactic constructs of the English Grammar: adjectival (or relative) clauses, adverbial clauses, coordinate clauses, subordinate clauses, correlated clauses, participial phrases, appositive phrases and voice. Passive voice is changed into active voice, while the remaining simplifications split a complex sentence into two (or more) with a subsequent decision about sentence order based on discourse organization.

Williams and Reiter [15, 29] use psycholinguistic findings on readability as a basis to their easy-to-read text generation system: short, common words are easier to read; short sentences are more readable; discourse connectives improve comprehension; cognitive load for poor readers in working out ellipses can be higher; some repetition and redundancy might actually turn out to be beneficial. They also use corpus analysis to search for cue phrases preferences and positions, and order of text spans, for instance.

The Coh-Metrix 2.0 tool [34] measures syntactic complexity. One of its metrics is very interesting: the number of words that appear before the main verb of the main clause in the sentences of a text. Sentences that have many words and subordinate clauses before the main verb demand a large amount of working memory.

3.3 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 of important pieces of information (bold, italic and larger 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 want to measure how the 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 Brazilian Portuguese. They belong to different genres and are available on the Web:

(1.) Corpus *Ao Encontro da Lei* (hereafter *Enc*), described in Section 2;

(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 also 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 analyze several features of a corpus, available at http://www.nilc.icmc.usp.br/AIC/). In order to compute the number of simple words in each corpus, we used the list of common words for Portuguese mentioned in Section 3.1. The discourse markers counted were those identified by [4] for Brazilian Portuguese. 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.

Table 1. Simple statistics of 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

We can see that all the 6 corpora of simple account texts have fewer words per sentence than the *FSP* corpus, that is, the non-simple account text. They also contain more common words. The The ANOVA statistical test showed that the difference between simple accounts and normal texts is significant with p-value < 0.05.

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 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 – lists with bullets or numbers for several aspects 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.

The ANOVA statistical test showed that the difference between simple accounts and normal texts for apposition and passive sentences are not strongly significant (p-values were 0.18 and 0.25, respectively).

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 shows that the simple account corpora, except those aiming children, contain proportionally more sentences with only one or two clauses (1-clause sentences and 2-clause sentences) than *FSP* corpus, that is, *FSP* seems indeed to contain more complex syntactic constructions.

This finding regarding to the simple accounts aimed to children was curious for several reasons. For example, *ZH* 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 was reduced.

ANOVA statistical test showed that the difference between simple accounts and normal texts for initial subordinate clauses and 2 to 5-clause sentences are not significant as expected (p-values were 0.34, 0.72, 0.61, 0.52 and 0.09, respectively).

Analyzing discourse markers, we noticed that there are more 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).

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 [37] 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 and other rules.

As shown in Table 4, 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 simplifications operations. The constructs are: (1) apposition, (2) relative clauses, (3) subordinate clauses, (4) coordinate clauses, (5) sentences with non-finite verbs, and (6) passive voice. 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: shorten sentences; keep the subject-verb-object together; avoid embedded sentence between parentheses, commas or dashes.

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 (%)	Subordi-nating 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

Table 4. Simplification operations in the manual

Construct	*Op.*	*Order of clauses*	*Cue phrase*	*Comments*
1.Enumerative appositive	e			Used to list items in simple accounts
1.Embedded appositive (app.)	a	Original/ App.		Appositive: Subject is the head of original + to be in present tense + apposition
2.Non-restrictive relative clause	a	Original/ Relative		Relative: Subject is the head of original + relative
2. Restrictive relative clause	a	Relative/ Original		Relative: Subject is the head of original + relative
3.Causal/Reason subordinate clause	a, b, d	Sub/ Main	*With this*	To keep the ordering cause, result
3.Comparative subordinate clause	a, b	Main/ Sub	*Also*	Rule for *such ... as*, *so ... as* markers
	e			Rule for the others markers or short sentences
3.Concessive subordinate clause	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
3.Conditional subordinate clause	e			Pervasive use in simple accounts
3. Result subordinate clause	a, b	Main/ Sub	*Thus*	May need some changes in verb
3.Final/Purpose subordinate clause	a, b	Main/ Sub	*The goal is*	
3.Proportional subordinate clause	e			Sub. clause frequently appears without a verb
3.Confirmative subordinate clause	a, b, d	Sub/ Main	*Confirms that*	May need some changes in verb
3.Temporal subordinate clause	a, b		*Then*	Rule for the markers: after that, as soon as
4.Asyndetic coordinate clause	a	Keep order		New sentences: Subjects are the head of the original subject
4.Additive coordinate clause	a	Keep order	Keep marker	Keep marker; it appears in the beginning of the new sentence
4.Adversative coordinate clause	a, b	Keep order	*But*	
4.Correlated coordinate clause	a, b	Keep order	*Also*	Original markers disappear
4.Result coordinate clause	a, b	Keep order	*As a result*	
4.Reason coordinate clause	a, b	Keep order	*This happens because*	May need some changes in verb
5.Non-finite verbs	e			Used to shorten sentences
6.Passive voice	c			

Table 4 shows the construct, the simplification operations to be applied, the suggested order of the clauses, and the cue phrase(s) (translated into English), if they apply. 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 (7th line in Table 4), 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.

4.1 Natural and Strong Simplifications

In the PorSimples project we are addressing TS to allow poor literacy people to have easier access to information. As described in Section 1, readers with literacy at basic level may need different type of help from those with literacy at rudimentary level, children learning to read, and people with cognitive disabilities. In fact, several researchers relate the capabilities and performance of the working memory with reading levels (see [8], for example). Beginning and poor readers tend to overload their working memory while trying to recognize words, which is considered to be a low level ability. Several studies have also shown that splitting complex sentences in shorter sentences (which is one of the many possible syntactic simplifications that can be done) results in the reduction of information in the working memory. On the other hand, in

PorSimples we also want to help poor literacy people 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 simple texts. Although simplification is an educational action that teachers perform on a daily basis, this action must be well balanced to improve students' learning skills.

In order to pursuit such 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 newspaper articles in Portuguese: from all the operations related in Table 4, plus lexical simplification and dropping sentences (or parts of them) which are redundant (not covered by the syntactic simplification guidelines), sentence splitting was the only one operation used with parsimony. This exercise helped us to propose these two levels of simplification that can be tuned according to user's needs. The *natural* simplification subset includes lexical simplification, dropping parts of the text, changing sentences to keep subject-verb-object order, changing discourse markers by simpler and/or more frequent ones, changing passive by active voice, inverting clause order, and non-simplification. *Strong* simplifications involve splitting sentences and changing discourse markers by simpler and/or more frequent ones, and inverting clause order.

In the following, we use the first paragraph of an article of the FSP newspaper (section Brazil, 2005), translated into English here, to illustrate the use of natural simplifications to produce a simplified text which is afterwards further simplified by using strong simplification operations:

Em entrevista coletiva convocada para responder às acusações de ter cobrado propina na Prefeitura de Ribeirão Preto, o ministro Antonio Palocci Filho (Fazenda) disse que colocou o cargo à disposição, mas, por orientação do presidente Luiz Inácio Lula da Silva, permanecerá no governo. Ressalvou insistentemente, porém, que não é "insubstituível".

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 office. He strongly stated, however, that no person is "irreplaceable".

After dropping the underlined parts above, since they appear further in the text, we have the following text:

Em entrevista coletiva, o ministro Antonio Palocci (Fazenda) disse que colocou o cargo à disposição, mas, por orientação do presidente Lula, permanecerá no governo. Ressalvou insistentemente, porém, que não é "insubstituível".

After lexical simplification of the underlined parts of the , we have the new text below which still needs some rewriting in the last sentence (to move the discourse marker "porém" (however) to the front):

Em entrevista coletiva, o ministro Antonio Palocci (Fazenda) disse que pode deixar o cargo de ministro, mas, presidente Lula orientou, continuar no governo. Insistiu, porém, que não é "insubstituível".

Finally, after changing the sentence to subject-verb-object ordering and moving the discourse marker, we have a *natural* simplified text, but it still has three clauses in the first sentence:

O Ministro Antonio Palocci (Fazenda) disse em entrevista coletiva que pode deixar o cargo, mas que o presidente Lula o orientou a continuar no governo. Porém, Palocci insistiu que não é "insubstituível".

Using the splitting operation in two constructs, embedded appositive "(*Fazenda*)" and adversative coordinate clause (starting with the marker "*mas*" (but)), we produce the final *strong* simplified text, with four sentences:

O Ministro Antonio Palocci é ministro da Fazenda. Antonio Palocci disse em entrevista coletiva que pode deixar o cargo. Mas ele disse que o presidente Lula o orientou a continuar no governo. Porém, Palocci insistiu que não é "insubstituível".

5. FINAL REMARKS AND FUTURE WORK

We presented in this paper the first steps towards producing TS systems for Brazilian Portuguese texts under the PorSimples project, which aims to facilitate information access by poor literacy people. 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 yielded the first syntactic simplification manual for Brazilian Portuguese and the grouping of the simplification operations in two subsets: natural and strong simplifications. The manual will serve as a basis for annotating corpora and for producing automatic TS systems, the immediate future work we foresee in the project. Initial work on a text simplification supporting tool can be found in [38].

We intend to evaluate our TS systems and the simplification operations in the manual by conducting experiments with poor readers of varied levels and different reading disability causes. We believe that it is possible to identify simplification operation groups tailored to different readers.

6. ACKNOWLEDGMENTS

We thank FAPESP and Microsoft Research for supporting the PorSimples project. The authors would like to thank Helena M. Caseli, Carol Scarton, Tiago Pereira, Rachel Aires, Amanda Rocha, Arnaldo Candido Jr. and Paulo Margarido for their valuable collaboration in the PorSimples project.

7. REFERENCES

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

[2] Rino, L.H.M., Pardo, T.A.S., Silla Jr., C.N., Kaestner, C.A., Pombo, M. 2004. A Comparison of Automatic Summarization Systems for Brazilian Portuguese Texts. SBIA 2004, Lecture Notes in Artificial Inteligence. 3171, Springer-Verlag, Berlin Heidelberg New York, 235-244.

[3] Feltrim, V., Pelizzoni, J.M., Teufel, S., Nunes, M.G.V., Aluísio, S.M. 2004. Applying Argumentative Zoning in an Automatic Critiquer of Academic Writing. SBIA 2004, Lecture Notes in Artificial Inteligence. 3171, Springer-Verlag, Berlin Heidelberg New York, 1-10.

[4] Pardo, T.A.S., Nunes, M.G.V. 2006. Review and Evaluation of DiZer - An Automatic Discourse Analyzer for Brazilian Portuguese. PROPOR 2006, Lecture Notes in Computer Science. 3960, Springer-Verlag, Berlin Heidelberg New York, 180-189.

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

[6] Max, A. 2006. Writing for Language-impaired Readers. In Proceedings of Seventh International Conference on Intelligent Text Processing and Computational Linguistics (Mexico City, Mexico, February 19-25, 2006). CICLing 2006. Springer-Verlag, Berlin Heidelberg New York, 567-570.

[7] Petersen, S. E., Ostendorf, M.: Text Simplification for Language Learners: A Corpus Analysis. 2007. In Proceedings of the Speech and Language Technology for Education Workshop (Pennsylvania, USA, October 1-3, 2007). SLaTE-2007. Carnegie Mellon University and ISCA Archive, http://www.isca-speech.org/archive/slate_2007. 69-72.

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

[9] Siddharthan, A. 2002. An Architecture for a Text Simplification System. In Proceedings of the Language Engineering Conference (Hyderabad, India, December 13-15, 2002). IEEE Computer Society 2002, 64-71.

[10] Klebanov, B., Knight, K., Marcu, D. 2004. Text Simplification for Information-Seeking Applications. On the Move to Meaningful Internet Systems. Lecture Notes in Computer Science. 3290, Springer-Verlag, Berlin Heidelberg New York, 735-747.

[11] Devlin, S. and Unthank, G. 2006. Helping aphasic people process online information. In Proceedings of the ACM SIGACCESS 2006, Conference on Computers and Accessibility (Portland, Oregon, USA, October 23-25, 2006). ASSETS 2006. New York: ACM, 225-226.

[12] Chandrasekar R., Doran C. and Srinivas, B. 1996. Motivations and Methods for Text Simplification. In Proceedings of the 16th International Conference on Computational Linguistics (Center for Sprogteknologi, Copenhagen, Denmark, August 5-9, 1996). COLING 1996, 1041-1044.

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

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

[15] Williams, S., Reiter, E. 2003. A corpus analysis of discourse relations for Natural Language Generation. In Proceedings of the Corpus Linguistics 2003 (Lancaster, England, March 28 – 31, 2003), CL2003, 899-908.

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

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

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

[19] Freire, A.P., Fortes, R.P.M, Paiva, D.M.B., Turine, M.A.S., 2007. Using Screen Readers to Reinforce Web Accessibility Education. In Proceedings of the 12th ACM Annual SIGCSE Conference on Innovation and Technology in Computer Science Education (Dundee, Scotland, June 25-27, 2007). ITiCSE 2007. ACM Press, New York, NY, 82-86.

[20] Freire, A.P., Fortes, R.P.M. 2005. Automatic accessibility evaluation of dynamic web pages generated through XSLT. In Proceedings of the International Cross-Disciplinary Workshop on Web Accessibility. (Chiba, Japan, May 10-14, 2005). W4A 2005. ACM Press, New York, NY, 81-84.

[21] Freire, A P., Goularte, Fortes, R. P. M. 2007. Techniques for Developing More Accessible Web Applications: a Survey Towards a Process Classification. In Proceedings of 25th ACM International Conference on Design of Communication. (El Paso, Texas, EUA, October 22-24, 2007). SIGDOC 2007. ACM Press, New York, NY, 162-169.

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

[23] Inui, K.; Fujita, A., Takahashi, T., Iida, R., Iwakura, T. 2003. Text simplification for reading assistance: a project note. In Proceedings of the Second International Workshop on Paraphrasing (Sapporo, Japan, July 11, 2003). IWP2003. Association for Computational Linguistics, Morristown, NJ, USA, 9-16.

[24] Daelemans, W., Hothker, A., Sang, E.T.K. 2004. Automatic Sentence Simplification for Subtitling in Dutch and English. In Proceedings of the 4th International Conference on Language Resources and Evaluation (Lisbon, Portugal, May 26-28, 2004), LREC 2004. ELRA Paris, France, 1045-1048.

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

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

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

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

[29] Williams S., Reiter E. 2008. Generating basic skills reports for low-skilled readers, Natural Language Engineering, First View article, (Apr. 2008), 1-31. Published online by Cambridge University Press 24 Apr 2008.

[30] Williams S., Reiter E. 2005. Generating Readable Texts for Readers with Low Basic Skills. In Proceedings of the 10th European Workshop on Natural Language Generation (Aberdeen, Scotland, August 8-10, 2005). ENLG-2005,

Association for Computational Linguistics, Morristown, NJ, USA, 140-147.

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

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

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

[34] 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, 36, 193-202.

[35] Muniz, M., Paulovich, F. V., Minghim, R., Infante, K., Muniz, F., Vieira, R., Aluísio, S. 2007. Taming the tiger topic: an XCES compliant corpus Portal to generate subcorpus based on automatic text topic identification. In Proceedings of the Corpus Linguistics 2007 (University of Birmingham, July 27-30, 2007). CL 2007.

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

[37] Specia, L.; Aluisio, S.M.; Pardo, T.A.S. 2008. Manual de Simplificação Sintática para o Português. Technical Report NILC-TR-08-06. São Carlos-SP. http://www.nilc.icmc.usp.br/nilc/publications.htm#TechnicalReports

[38] Aluísio, S.M.; Specia, L.; Pardo, T.A.S.; Maziero, E.G.; Caseli, H.M.; Fortes, R.P.M. “A Corpus Analysis of Simple Account Texts and the Proposal of Simplification Strategies: First Steps towards Text Simplification Systems”, Proceedings of the 26th ACM International Conference on Design of Communication, 2008, in press.