Fostering Digital Inclusion and Accessibility: The PorSimples project for Simplification of Portuguese Texts

Sandra Maria Aluísio and Caroline Gasperin
Department of Computer Sciences, University of São Paulo
Av. Trabalhador São-Carlense, 400. 13560-970 - São Carlos/SP, Brazil
{sandra,cgasperin}@icmc.usp.br

Abstract

In this paper we present the PorSimples project, whose aim is to develop text adaptations tools for Brazilian Portuguese. The tools developed cater for both people at poor literacy levels and authors that want to produce texts for this audience. Here we describe the tools and resources developed over two years of this project and point directions for future work and collaboration. Since Portuguese and Spanish have many aspects in common, we believe our main point for collaboration lies in transferring our knowledge and experience to researches willing to developed simplification and elaboration tools for Spanish.

1 Introduction

In Brazil, according to the index used to measure the literacy level of the population (INAF - National Indicator of Functional Literacy) (INAF, 2007), only 28% of the population is classified as literate at the advanced level, while 65% of the population face difficulties in activities involving reading and comprehension depending on text length and complexity; therefore, their access to textual media is limited. The latter ones belong to the so-called *rudimentary* and *basic* literacy levels. These people are only able to find explicit information in short texts (rudimentary level) and also process slightly longer texts and make simple inferences (basic level).

The production of texts with different lengths and complexities can be addressed by the task of Text Adaptation (TA), a very well known practice in educational settings. Young (1999) and Burstein (2009) mention two different techniques for TA: *Text Simplification* and *Text Elaboration*.

The first can be defined as any task that reduces the lexical or syntactic complexity of a text, while trying to preserve meaning and information. Text Simplification can be subdivided into Syntactic Simplification, Lexical Simplification, Automatic Summarization, and other techniques.

As to Text Elaboration, it aims at clarifying and explaining information and making connections explicit in a text, for example, providing short definitions or synonyms for words known to only a few speakers of a language.

The PorSimples project[1] (Simplification of Portuguese Text for Digital Inclusion and Accessibility) (Aluisio et al, 2008a) started in November 2007 and will finish in April 2010. It aims at developing technologies to make access to information easier for low-literacy individuals, and possibly for people with other kinds of reading disabilities, by means of Automatic Summarization, Lexical Simplification, Syntactic Simplification, and Text Elaboration. More specifically, the goal is to help these readers to process documents available on the web. Additionally, it could help children learning to read texts of different genres, adults being alphabetized, hearing-impaired people who communicate to each other using sign languages and people undertaking Distance Education, in which text intelligibility is of great importance.

The focus is on texts published in government sites or by relevant news agencies, both of impor-

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

Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas, pages 46–53, Los Angeles, California, June 2010. ©2010 Association for Computational Linguistics

tance 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.

In the project we have developed resources in Portuguese for research on text simplification, text simplification technology for Portuguese, and currently we are developing and adapting resources and technologies for text elaboration. We have also built applications that make the developed technology available to the public. In the Sections 2 to 4 we describe all these outcomes of the project.

We intend to foster a new interdisciplinary research area to study written text comprehension problems via the research on readability assessment, text simplification and elaboration once PorSimples ends. In Section 5 we describe future work, and in Section 6 we outline potential points for collaboration with researchers from Brazil and the rest of the Americas.

2 Resources

In order to understand the task of text simplification in Portuguese and to build training and evaluation data for the systems developed in the project, we have created a set of resources that formed the basis of PorSimples. Moreover, we are currently working on building resources for text elaboration. Below we describe these resources.

2.1 Manual for Syntactic Simplification in Portuguese

We have created a Manual for Syntactic Simplification for Portuguese (Specia et al., 2008). This manual recommends how particular syntactic phenomena should be simplified. It is based on a careful study of the Brazilian Portuguese grammar, of simplification systems developed for English (for example, (Siddharthan, 2003)), and on the Plain Language initiative[2] (Aluisio et al., 2008b).

The manual was the basis for the development of our rule-based system for syntactic simplification described in Section 3.2.

2.2 Corpora of Simple and Simplified Texts

We have built 9 corpora within 2 different genres (general news and popular science articles). Our first corpus is composed of general news articles from the Brazilian newspaper Zero Hora (ZH original). We had these articles manually simplified by a linguist, specialized in text simplification, according to the two levels of simplification proposed in PorSimples, natural (ZH natural) and strong (ZH strong). The Zero Hora newspaper also provides along its articles a simple version of them targeting children from 7 to 11 years old; this section is called *Para seu Filho Ler* (ZH PFSL) and our corpus from this section contains simple articles corresponding to the articles in the ZH original corpus plus additional ones.

Popular science articles compose our next set of corpora. We compiled a corpus of these articles from the *Caderno Ciência* issue of the Brazilian newspaper Folha de São Paulo, a leading newspaper in Brazil (CC original). We also had this corpus manually simplified according to the natural (CC natural) and strong (CC strong) levels. We also collected texts from a popular science magazine called *Ciência Hoje* (CH) and from its version aimed at children from 12-15, called *Ciência Hoje Crianca* (CHC). Table 1 shows a few statistics from these corpora.

2.3 Dictionary of Simple Words

While for English some lexical resources that help to identify difficult words using psycholinguistic measures are available, such as the MRC Psycholinguistic Database[3], no such resources exist for Portuguese. In PorSimples, we have compiled a dictionary of simple words composed by words that are common to youngsters (from Biderman (2005)), a list of frequent words from news texts for children and nationwide newspapers and a list of concrete words (from Janczura et. al (2007)).

Corpus	Art.	Sent.	Words	Avg. words per text (std. deviation)	Avg. words p. sentence
ZH original	104	2184	46190	444.1 (133.7)	21.1
ZH natural	104	3234	47296	454.7 (134.2)	14.6
ZH strong	104	3668	47938	460.9 (137.5)	13.0
ZH PSFL	166	1224	22148	133.4 (48.6)	18.0
CC original	50	882	20263	405.2 (175.6)	22.9
CC natural	50	975	19603	392.0 (176.0)	20.1
CC strong	50	1454	20518	410.3 (169.6)	14.1
CH	130	3624	95866	737.4 (226.1)	26.4
CHC	127	3282	65124	512.7 (185.3)	19.8

Table 1. Corpus statistics.

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

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

This dictionary is being used in applications described in Section 4, such as SIMPLIFICA and the Simplification Annotation Editor.

3 Simplification & Elaboration technology

3.1 Lexical Simplification

Lexical simplification consists on replacing complex words by simpler words.

The first step of lexical simplification consists of tokenizing the original text and selecting the words that are considered complex. In order to judge a word as complex or not, we use the dictionaries of simple words described in Section 2.3.

The lexical simplification system also uses the Unitex-PB dictionary[4] for finding the lemma of the words in the text, so that it is possible to look for it in the simple words dictionaries. The problem of looking for a lemma directly in a dictionary is that there are ambiguous words and we are not able to deal with different word senses. For dealing with part-of-speech (POS) ambiguity, we use the MXPOST POS tagger[5] trained over NILC tagset[6].

Among the words that were selected as complex, the ones that are not proper nouns, prepositions and numerals are processed: their POS tags are used to look for their lemmas in the dictionaries. As the tagger has not a 100% precision and some words may not be in the dictionary, we look for the lemma only (without the tag) when we are not able to find the lemma-tag combination in the dictionary. Still, if we are not able to find the word, the lexical simplification module assumes that the word is complex and marks it for simplification.

The last step of the process consists in providing simpler synonyms for the complex words. For this task, we use the thesauri for Portuguese TeP 2.0[7] and the lexical ontology for Portuguese PAPEL[8]. This task is carried out when the user clicks on a marked word, which triggers a search in the thesauri for synonyms that are also present in the common words dictionary. If simpler words are found, they are sorted from the simpler to the more complex. To determine this order, we used Google API to search each word in the web: we assume that the higher a word frequency, the simpler it is. Automatic word sense disambiguation is left for future work. In PorSimples, we aim to use Textual Entailment (Dagan et al., 2005) as a method for gathering resources for lexical simplification.

3.2 Syntactic Simplification

Syntactic simplification is accomplished by a rule-based system, which comprises seven operations that are applied sentence-by-sentence to a text in order to make its syntactic structure simpler.

Our rule-based text simplification system is based on the manual for Brazilian Portuguese syntactic simplification described in Section 2.1. According to this manual, simplification operations should be applied when any of the 22 linguistic phenomena covered by our system (see Candido et al. (2009) for details) is detected. Our system treats appositive, relative, coordinate and subordinate clauses, which have already been addressed by previous work on text simplification (Siddharthan, 2003). Additionally, we treat passive voice, sentences in an order other than Subject-Verb-Object (SVO), and long adverbial phrases. The simplification operations to treat these phenomena are: split sentence, change particular discourse markers by simpler ones, change passive to active voice, invert the order of clauses, convert to subject-verb-object ordering, and move long adverbial phrases.

Each sentence is parsed in order to identify syntactic phenomena for simplification and to segment the sentence into portions that will be handled by the operations. We use the parser PALAVRAS (Bick, 2000) for Portuguese. Gasperin et al. (2010) present the evaluation of the performance of our syntactic simplification system.

Since our syntactic simplifications are conservative, the simplified texts become longer than the original due to sentence splitting. We acknowledge that low-literacy readers prefer short texts; this is why we use summarization before applying simplification in FACILITA (see (Watanabe et al., 2009)). In the future we aim to provide summarization also within SIMPLIFICA. These two applications are described in Section 4.

3.3 Natural and Strong Simplification

To attend the needs of people with different levels of literacy, PorSimples propose two types of sim-

[4] http://www.nilc.icmc.usp.br/nilc/projects/unitex-pb/web/dicionarios.html

[5] http://sites.google.com/site/adwaitratnaparkhi/home

[6] www.nilc.icmc.usp.br/nilc/TagSet/ManualEtiquetagem.htm

[7] http://www.nilc.icmc.usp.br/tep2/

[8] http://www.linguateca.pt/PAPEL/

plification: natural and strong. The first is aimed at people with a basic literacy level and the second, rudimentary level. The difference between these two is the degree of application of simplification operations to the sentences. For strong simplification we apply the syntactic simplification process to all complex phenomena found in the sentence in order to make the sentence as simple as possible, while for natural simplification the simplification operations are applied only when the resulting text remains "natural", considering the overall complexity of the sentence. This naturalness is based on a group of factors which are difficult to define using hand-crafted rules, and we intend to learn them from examples of natural simplifications.

We developed a corpus-based approach for selecting sentences that require simplification. Based on parallel corpora of original and natural simplified texts (ZH original, ZH natural, CC original, CC natural), we apply a binary classifier to decide in which circumstances a sentence should be split or not so that the resulting simplified text is natural and not over simplified. Sentence splitting is the most important and most frequent syntactic simplification operation, and it can be seen as a key distinctive feature between natural and strong simplification. We described this system in detail in (Gasperin et al., 2009).

Our feature set contains 209 features, including superficial, morphological, syntactic and discourse-related features. We did several feature selection experiments to determine the optimal set of features. As classification algorithm we use Weka's[9] SMO implementation of Support Vector Machines (SVM). The ZH corpus contains 728 examples of the splitting operation and 1328 examples of non-split sentences, and the CC corpus contains 59 positive and 510 negatives examples. The classifier's average performance scores (optimal feature set, both corpora as training data, and cross-validation) are 80.5% precision and 80.7% recall.

3.4 Readability Assessment

We developed a readability assessment system that can predict the complexity level of a text, which corresponds to the literacy level expected from the target reader: *rudimentary, basic* or *advanced*.

We have adopted a machine-learning classifier to identify the level of the input text; we use the Support Vector Machines implementation from Weka toolkit (SMO). We have used 7 of our corpora presented in Section 2.2 (all but the ones with texts written for children) to train the classifier.

Our feature set is composed by cognitively-motivated features derived from the Coh-Metrix-PORT tool[10], which is an adaptation for Brazilian Portuguese of Coh-Metrix 2.0 (free version of Coh-Metrix (Graesser et al, 2004)) also developed in the context of the PorSimples project. Coh-Metrix-PORT implements the metrics in Table 2.

We also included seven new metrics to Coh-Metrix-PORT: average verb, noun, adjective and adverb ambiguity, incidence of high-level constituents, content words and functional words.

Categories	Subcategories	Metrics
Shallow Readability metric	-	Flesch Reading Ease index for Portuguese.
Words and textual information	Basic counts	Number of words, sentences, paragraphs, words per sentence, sentences per paragraph, syllables per word, incidence of verbs, nouns, adjectives and adverbs.
	Frequencies	Raw frequencies of content words and minimum frequency of content words.
	Hyperonymy	Average number of hypernyms of verbs.
Syntactic information	Constituents	Incidence of nominal phrases, modifiers per noun phrase and words preceding main verbs.
	Pronouns, Types and Tokens	Incidence of personal pronouns, number of pronouns per noun phrase, types and tokens.
	Connectives	Number of connectives, number of positive and negative additive connectives, causal / temporal / logical positive and negative connectives.
Logical operators	-	Incidence of the particles "e" (and), "ou" (or), "se" (if), incidence of negation and logical operators.

Table 2. Metrics of Coh-Metrix-PORT.

We measured the performance of the classifier on identifying the levels of the input texts by a

[9] http://www.cs.waikato.ac.nz/ml/weka/

[10] http://caravelas.icmc.usp.br:3000/

cross-validation experiment. We trained the classifier on our 7 corpora and reached 90% F-measure on identifying texts at advanced level, 48% at basic level, and 73% at rudimentary level.

3.5 Semantic Role Labeling: Understanding Sense Relations between Verb and Arguments

To attend the goal of eliciting sense relations between verbs and their arguments through the exhibition of question words such as *who, what, which, when, where, why, how, how much, how many, how long, how often* and *what for*, we are specifying a new annotation task that assigns these wh-question labels to verbal arguments in a corpus of simplified texts in Portuguese. The aim is to provide a training corpus for machine learning, aiming at automatic assignment of wh-questions (Duran et al., 2010a; Duran et al., 2010b).

The annotation task involves recognizing segments that constitute answers to questions made to the verbs. Each segment should suitably answer the wh-question label. For example, in the sentence "João acordou às 6 horas da manhã." (John woke up at 6 in the morning.), two questions come up naturally in relation to the verb "acordar" (wake up): 1) *Who woke up?* and 2) *When?*.

Linking the verb and its arguments through wh-questions is a process that requires text understanding. This is a skill that the target audience of this project is weak at. In Figure 1 we show the link between the verb and its arguments (which can be subject, direct object, indirect object, time or location adverbial phrases, and also named entities).

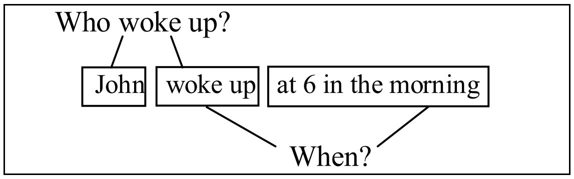

Figure 1. Assigning wh-question labels to arguments.

The corpus chosen for this work consists of the strong simplified version of 154 texts extracted from general news and popular science articles (ZH strong and CC strong) which were described in Section 2.2.

Results of such a semantic layer of annotation may be used, in addition, to identify adjunct semantic roles and multi-word expressions with specific adverbial syntactic roles. This training corpus, as well as the automatic labeling tool, an "answer-questioning" system, will be made publicly available at PorSimples site. Besides helping poor-literacy readers, the assignment of wh-questions will be used in the near future to map adjunct semantic roles (ArgMs of Propbank (Palmer et al., 2005)) in a project to build the PropBank.Br for Portuguese language. One may also take profit of this automatic tool and its training corpus to improve its opposite, question-answering systems.

4 Applications

The text simplification and elaboration technologies developed in the context of the project are available by means of three systems aimed to distinct users:

- An authoring system, called SIMPLIFICA[11], to help authors to produce simplified texts targeting people with low literacy levels,
- An assistive technology system, called FACILITA[12], which explores the tasks of summarization and simplification to allow poor literate people to read Web content, and
- A web content adaptation tool, named Educational FACILITA, for assisting low-literacy readers to perform detailed reading. It exhibits questions that clarify the semantic relations linking verbs to their arguments, highlighting the associations amongst the main ideas of the texts, named entities, and perform lexical elaboration.

In the following subsections we detail these and other systems developed in the project.

4.1 SIMPLIFICA Authoring Tool

SIMLIFICA is a web-based WYSIWYG editor, based on TinyMCE web editor[13]. The user inputs a text in the editor and customizes the simplification settings, where he/she can choose: (i) strong simplification, where all the complex syntactic phenomena (see details in Section 3.2) are treated for each sentence, or customized simplification, where the user chooses one or more syntactic simplification phenomena to be treated for each sentence, and (ii) one or more thesauri to be used in the syntactic and lexical simplification processes. Then

[11] http://www.nilc.icmc.usp.br/porsimples/simplifica/

[12] http://vinho.intermidia.icmc.usp.br:3001/facilita/

[13] http://tinymce.moxiecode.com/

the user activates the readability assessment module to predict the complexity level of a text. This module maps the text to one of the three levels of literacy defined by INAF: *rudimentary, basic* or *advanced*. According to the resulting readability level the user can trigger the lexical and/or syntactic simplifications modules, revise the automatic simplification and restart the cycle by checking the readability level of the current version of the text.

4.2 FACILITA

FACILITA is a browser plug-in that aims to facilitate the reading of online content by poor literate people. It includes separate modules for text summarization and text simplification. The user can select a text on any website and call FACILITA to summarize and simplify this text. The system is described in details in Watanabe et al. (2009).

The text summarization module aims to extract only the most important information from a text. It relies on the EPC-P technique (extraction of keywords per pattern), which checks the presence of keywords in the sentences: sentences that contain keywords are retained for the final summary. The summarization system is reported in Margarido et al. (2008).

The text simplification module follows the syntactic simplification framework described in Section 3.2. We have chosen to run the summarization process first and then proceed to the simplification of the summarized text since simplification increases text length.

4.3 Educational FACILITA

Educational FACILITA[14] is a Web application aimed at assisting users in understanding textual content available on the Web. Currently, it explores the NLP tasks of lexical elaboration and named entity labeling to assist poor literacy readers having access to web content. It is described in Watanabe et al. (2010).

Lexical Elaboration consists of mechanisms that present users with synonymous or short definitions for words, which are classified as unusual or difficult to be understood by the users. This process relies on the framework developed for lexical simplification described in Section 3.1.

Named-entity labeling consists of displaying additional and complementary semantic and descriptive information about named entities that are contained on the Web sites text. The descriptions are extracted from Wikipedia.

It is expected that these additional information presented in the text by the proposed approach would help users better understand websites' textual content and allow users to learn the meaning of new or unusual words/expressions.

4.4 Simplification Annotation Editor

This editor[15] was created to support the manual simplification of texts for the creation of our corpus of simplified texts. It records and labels all the operations made by the annotator and encode texts using a new XCES[16]-based schema for linking the original-simplified information. XCES has been used in projects involving both only one language, e.g. American National Corpus (ANC)[17] (English) and PLN-BR[18] (Brazilian Portuguese); and multiple languages as parallel data, e.g.: CroCo[19] (English-German). However, to our knowledge, PorSimples is the first project to use XCES to encode original-simplified parallel texts and also the simplification operations. Two annotation layers have been added to the traditional stand-off annotation layers in order to store the information related to simplification (Caseli et al., 2009).

4.5 Portal of Parallel Corpora

The portal[20] allows for online querying and download of our corpora of simplified texts. The queries can include information about syntactic constructions, simplification operations, etc.

5 Future Work

Our main area for future work lies on the evaluation of the simplified texts resulting from our systems with the end user, that is, people at low literacy levels. We are carrying out a large-scale study with readers who fit in the rudimentary and basic literacy levels to verify whether syntactic and lexi-

[14] http://vinho.intermidia.icmc.usp.br/watinha/Educational-Facilita/

[15] http://caravelas.icmc.usp.br/anotador

[16] http://www.w3.org/XML/

[17] http://americannationalcorpus.org

[18] http://www.nilc.icmc.usp.br/plnbr

[19] http://fr46.uni-saarland.de/croco/index_en.html

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

cal simplification indeed contribute to the understanding of Portuguese texts. We are applying reading comprehension tests with original texts (control group) and manually simplified texts at strong level. However we still need to assess the impact of automatic lexical and syntactic simplification and text elaboration on the understanding of a text by the target user of our applications.

We also intend to investigate how to balance simplification/elaboration and text length. We have shown that in our syntactic simplification approach it is usual to divide long sentences, which reduce sentence length but increase text length due to the repetition of the subject in the new sentences. On the other hand, in summarization-based Text Simplification, such as FACILITA's approach, text length is reduced, but relevant information can be lost, which may hinder text comprehensibility. Text Elaboration enhances text comprehensibility, but it always increases text length, since it inserts information and repetition to reinforce understanding and make explicit the connections between the parts of a text. Therefore, since we cannot achieve all the requisites at once there is a need to evaluate each aspect of our systems with the target users.

We also intend to improve the performance of our syntactic simplification approach by experimenting with different Portuguese syntactic parsers. Moreover, several methods of text elaboration are still under development and will be implemented and evaluated in this current year.

As future research, we aim to explore the impact of simplification on text entailment recognition systems. We believe simplification can facilitate the alignment of entailment pairs. In the opposite direction, text entailment or paraphrase identification may help us find word pairs for enriching the lexical resources used for lexical simplification.

6 Opportunities for Collaboration

Enhancing the accessibility of Portuguese and Spanish Web texts is of foremost importance to improve insertion of Latin America (LA) into the information society and to preserve the diverse cultures in LA. We believe several countries in LA present similar statistics to Brazil in relation to the number of people at low literacy levels. We see our experience in developing text simplification and elaboration tools for Portuguese as the major contribution that we can offer to other research groups in LA. We are interested in actively taking part in joint research projects that aim to create text simplification and elaboration tools for Spanish.

Since all resources that we have developed are language-dependent, they cannot be used directly for Spanish, but we foresee that due to similarities between Portuguese and Spanish a straightforward adaptation of solutions at the lexical and syntactical levels can be achieved with reasonable effort. We are willing to share the lessons learned during the PorSimples project and offer our expertise on selecting and creating the appropriate resources (e.g. corpora, dictionaries) and technology for text simplification and elaboration in order to create similar ones for Spanish.

The advances in text simplification and elaboration methods strongly depend on the availability of annotated corpora for several tasks: text simplification, text entailment, semantic role labeling, to name only a few. English has the major number of data resources in Natural Language Processing (NLP); Portuguese and Spanish are low-density languages. To solve this problem, we believe that there is a need for: (i) the development of a new area recently coined as Annotation Science; (ii) a centralized resource center to create, collect and distribute linguistic resources in LA.

We would appreciate collaboration with researchers in the USA in relation to readability assessment measures, such as those of Coh-Metrix (see Section 3.4), whose researchers already developed up to 500 measures. Only 60 of them are open to public access. Besides, the know-how needed to develop a proposition bank of Portuguese would be welcome since this involves lexical resources, such as a Verbnet[21], which do not exist for Portuguese. Other lexical resources such as the MRC Psycholinguistic Database, which help to identify difficult words using psycholinguistic measures, are also urgent for Portuguese since we have sparse projects dealing with several aspects of this database but no common project to unite them.

Brazilian research funding agencies, mainly CAPES[22], CNPq[23] and FAPESP[24], often release calls for projects with international collaboration; these could be a path to start the collaborative research suggested above.

[21] http://verbs.colorado.edu/~mpalmer/projects/verbnet.html

[22] http://www.capes.gov.br/

[23] http://www.cnpq.br/

[24] http://www.fapesp.br/

Acknowledgments

We thank FAPESP and Microsoft Research for supporting the PorSimples project.

References

Sandra Aluísio, Lucia Specia, Thiago Pardo, Erick Maziero and Renata Fortes. 2008a. Towards Brazilian Portuguese Automatic Text Simplification Systems. In: *Proceedings of The Eight ACM Symposium on Document Engineering (DocEng 2008)*, 240-248, São Paulo, Brazil.

Sandra Aluísio, Lucia Specia, Thiago Pardo, Erick Maziero, Helena de M. Caseli, Renata Fortes. 2008b. A Corpus Analysis of Simple Account Texts and the Proposal of Simplification Strategies: First Steps towards Text Simplification Systems In: *Proceedings of The 26th ACM Symposium on Design of Communication (SIGDOC 2008),* pp. 15-22.

Eckhard Bick. 2000. The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. PhD Thesis. Aarhus University.

Maria Teresa Biderman. 2005. DICIONÁRIO ILUSTRADO DE PORTUGUÊS. São Paulo, Editora Ática. 1ª. ed. São Paulo: Ática. (2005)

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

Arnaldo Candido Junior, Erick Maziero, Caroline Gasperin, Thiago Pardo, Lucia Specia and Sandra M. Aluisio. 2009. Supporting the Adaptation of Texts for Poor Literacy Readers: a Text Simplification Editor for Brazilian Portuguese. In the *Proceedings of the NAACL HLT Workshop on Innovative Use of NLP for Building Educational Applications*, pages 34–42, Boulder, Colorado, June 2009.

Helena Caseli, Tiago Pereira, Lucia Specia, Thiago Pardo, Caroline Gasperin and Sandra Aluísio. 2009. Building a Brazilian Portuguese parallel corpus of original and simplified texts. In Alexander Gelbukh (ed), *Advances in Computational Linguistics, Research in Computer Science*, vol 41, pp. 59-70. 10th Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2009).

Ido Dagan, Oren Glickman and Bernado Magnini. 2005. The PASCAL Recognising Textual Entailment Challenge. In: *Proceedings of The First PASCAL Recognising Textual Entailment Challenge* (RTE 1), [S.l.]: Springer, 2005. p. 1–8.

Magali Duran, Marcelo Amâncio and Sandra Aluísio. 2010a. Assigning wh-questions to verbal arguments in a corpus of simplified texts. Accepted for publication at *Propor* 2010 (http://www.inf.pucrs.br/~propor2010).

Magali Duran, Marcelo Amâncio and Sandra Aluísio. 2010b. Assigning Wh-Questions to Verbal Arguments: Annotation Tools Evaluation and Corpus Building. Accepted for publication in LREC 2010.

Caroline Gasperin, Lucia Specia, Tiago Pereira and Sandra Aluísio. 2009. Learning When to Simplify Sentences for Natural Text Simplification. In: *Proceedings of ENIA* 2009, 809-818.

Caroline Gasperin, Erick Masiero and Sandra M. Aluisio. 2010. Challenging choices for text simplification. Accepted for publication at *Propor* 2010 (http://www.inf.pucrs.br/~propor2010).

Arthur Graesser, Danielle McNamara, Max Louwerse and Zhiqiang Cai. 2004. Coh-Metrix: Analysis of text on cohesion and language. In: *Behavioral Research Methods, Instruments, and Computers*, 36, páginas 193-202.

INAF. 2007. Indicador de Alfabetismo Funcional INAF/Brasil - 2007. Available at http://www.acaoeducativa.org.br/portal/images/stories/pdfs/inaf2007.pdf

Gerson A Janczura, Goiara M Castilho, Nelson O Rocha, Terezinha de Jesus C. van Erven and Tin Po Huang. 2007. Normas de concretude para 909 palavras da língua portuguesa. *Psicologia: Teoria e Pesquisa* Abr-Jun 2007, Vol. 23 n. 2, pp. 195-204.

Martha Palmer, Daniel Gildea and Paul Kingsbury. 2005. The Proposition Bank: A Corpus Annotated with Semantic Roles, *Computational Linguistics Journal*, 31:1.

Advaith Siddharthan. 2003. Syntactic Simplification and Text Cohesion. PhD Thesis. University of Cambridge.

Lucia Specia, Sandra Aluisio and Tiago Pardo. 2008. Manual de Simplificação Sintática para o Português. Technical Report NILC-TR-08-06, 27 p. Junho 2008, São Carlos-SP.

Willian Watanabe, Arnaldo Candido Junior, Vinícius Uzêda, Renata Fortes, Tiago Pardo and Sandra Aluísio. 2009. Facilita: reading assistance for low-literacy readers. In: *Proceedings of the 27th ACM International Conference on Design of Communication. SIGDOC '09*. ACM, New York, NY, 29-36.

Willian Watanabe, Arnaldo Candido Junior, Marcelo Amancio, Matheus de Oliveira, Renata Fortes, Tiago Pardo, Renata Fortes, Sandra Aluísio. 2010. Adapting web content for low-literacy readers by using lexical elaboration and named entities labeling. Accepted for publication at W4A 2010 (http://www.w4a.info/).

Dolly J. Young. Linguistic simplification of SL reading material: effective instructional practice. *The Modern Language Journal*, 83(3):350–366, 1999.