Learning When to Simplify Sentences for Natural Text Simplification

Caroline Gasperin[1], **Lucia Specia**[1], **Tiago F. Pereira**[1], **Sandra M. Aluisio**[1]

[1]NILC - Núcleo Interinstitucional de Linguística Computacional
ICMC, Universidade de São Paulo
Av. Trabalhador São-Carlense, 400 - 13560-970 - São Carlos/SP, Brazil

{cgasperin,lspecia,sandra}@icmc.usp.br, tiagofrepereira@gmail.com

Abstract. *This paper introduces a corpus-based approach for selecting sentences that require simplification in the context of Brazilian Portuguese text simplification system. Based on a parallel corpus of original and simplified text versions, we apply a binary classifier to decide in which circumstances a sentence should or not be split – which is the most important syntactic simplification operation – so that the resulting simplified text is natural and not over simplified. Our classifier reaches 73.5% precision and 73.4% recall when selecting the sentences to be split or kept together.*

1. Introduction

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

The PorSimples project (*Simplificação Textual do Português para Inclusão e Acessibilidade Digital*[1]) aims at producing text simplification tools for promoting digital inclusion and accessibility for people with such levels of literacy, and possibly other kinds of reading disabilities. More specifically, the goal is to help these readers to process documents available on the web. Additionally, it could help children learning to read texts of different genres or adults being alphabetized. The focus is on texts published in government sites or by relevant news agencies, both expected to be of importance to a large audience with various literacy levels. The language of the texts is Brazilian Portuguese, for which there are no text simplification systems, to the best of our knowledge.

Text simplification has been exploited in other languages for helping poor literacy readers [Max 2006, Siddharthan 2003], bilingual readers [Petersen and Ostendorf 2007] and special kinds of readers such as aphasics [Devlin and Unthank 2006] and deaf people [Inui et al. 2003]. It has also been used for improving the accuracy of other natural language processing tasks [Chandrasekar and Srinivas 1997, Klebanov et al. 2004], like parsing and information extraction.

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

To attend the needs of people with different levels of literacy, we propose two types of simplification: *natural* and *strong*. The first type 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 a set of pre-defined simplification operations to make the sentence as simple as possible, while for natural simplification these operations are applied only when the resulting text remains "natural". 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.

The focus in this paper is on natural simplifications. We aim to learn from a corpus whether or not it is natural to simplify a given sentence. A corpus of natural simplifications from news articles was manually produced by a linguist, expert in text simplification, aiming to provide data for machine learning algorithms and insights in understanding under which conditions natural simplifications are produced. Analyzing the resulting simplified corpus, we observed that sentence splitting is the most frequent syntactic simplification operation used by the annotator when creating a natural simplified text. In this paper we therefore address the syntactic simplification problem of deciding whether to split a sentence or not. We apply a supervised machine-learning algorithm with a number of features for identifying sentences that should be split, and then pass these sentences on to a rule-based system [Jr. et al. 2009] that performs the actual simplification operations. Although such a system is also part of the PorSimples project, it is beyond the scope of this paper.

In the next section, we describe previous work on Text Simplification. In Section 3 we present the manual simplification process and the annotated corpora resulting from this process, giving more details about natural and strong simplifications. In Section 4 we describe the system for natural simplification that we have developed based on the annotated corpora, detail our experiments and present our results.

2. Related work

Existing text simplification systems can be compared along three axes: the type of system – rule-based or corpus-based –, the type of knowledge used to identify the need for simplification, and the goals of the system.

A few rule-based systems have been developed for text simplification [Chandrasekar et al. 1996, Siddharthan 2003], focusing on different readers (poor literate, aphasic, etc). These systems contain a set of manually created simplification rules that are applied to each sentence. These are usually based on parser structures and limited to certain simplification operations (like splitting relative clauses). Moreover, they do not cover different levels of simplification.

Corpus-based systems, on the other hand, can learn from corpus the relevant simplification operations and also the necessary degree of the simplification for a given task. The study that is closest to ours is that by [Petersen and Ostendorf 2007], but their goal is different: learning the rules governing the simplification in order to inform second language teachers. They adopt machine learning techniques in order to learn when to drop a sentence from the text and when to split a sentence. For splitting sentences, a C4.5 classifier is trained using 20 features (shallow, morphological and syntactic ones). An average

error rate of 29% is obtained in this classification task. The lengths of sentence and noun phrase were found to be the most important features.

[Chandrasekar and Srinivas 1997] also developed a corpus-based simplification system, where a rule-induction algorithm is applied on a corpus with chunks annotated using supertags – part-of-speech tags augmented by agreement and subcategorization information. Their goal, differently from ours, is to improve the performance of parsers and machine translation systems by providing them syntactically-simpler sentences as input.

To the best of our knowledge, none of the previous text simplification systems aims to provide varying degrees of simplification according to the user needs. Moreover, none of the existing systems addresses the language under consideration in this paper, Brazilian Portuguese, for which the need of text simplification is evident, given the high number of poor literacy readers, as mentioned in Section 1.

3. Corpus creation

The texts chosen to be annotated with its simplifications were extracted from two of the main Brazilian newspapers, *Zero Hora* and *Folha de São Paulo*. The *Zero Hora* texts are general news articles, chosen because they had a corresponding simplified version, also published in that newspaper, meant to be read by children. Therefore, this corpus can also be useful to evaluate the proposed simplifications against independently hand-crafted simplified versions. The *Folha de São Paulo* texts are from the *Caderno da Ciência* (Science section) and were selected because they present different characteristics from general news articles, as they comprise only science related topics. The goal was to collect corpora from different domains to validate our simplification techniques.

We developed a tool to assist human annotators in this inherently manual task – the Simplification Annotation Editor [2]. We also propose a new schema for representing the original-simplified information, based on the XCES standard[3]. The annotation tool and corpus encoding is detailed in [Caseli et al. 2009].

The Simplification Annotation Editor, besides facilitating the manual simplification process, records the simplification operations made by the annotator. The Editor has two modes to assist the human annotator: the Léxico and the Sintático modes. In the *Léxico* mode, the editor proposes changes in words and discourse markers by simpler and/or more frequent ones. The *Sintático* mode proposes syntactic operations based on syntactic clues provided by a parser for Portuguese [Bick 2000]. When the annotator selects an operation, it is recorded and the annotator can specify what has been changed in the simplified version.

Our set of lexical and syntactic simplification operations that can be applied to a sentence in the original text is the following: (1) non-simplification; (2) simple rewriting (replacing discourse markers or sets of words, like idioms or collocations) or (3) strong rewriting (any sort of free rewriting of sentence, as defined in [Petersen and Ostendorf 2007]); (4) putting the sentence in its canonical order (subject-verb-object); (5) putting the sentence in the active voice; (6) inverting the clause ordering; (7) splitting or (8) joining sentences; (9) dropping the sentence or (10) dropping parts of

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

[3] http://www.xml-ces.org

the sentence, and (11) lexical substitution. The lexical operations are (11) and (2), which consist of replacing words or longer expressions, respectively, found to be complex by simpler synonyms.

3.1. Natural versus strong simplification

Table 1 shows examples of an original text from an on-line Brazilian newspaper in (O), its natural simplification in (N) and its strong simplification in (S). The second and fourth sentences in (O) can be further simplified if split in shorter ones, as shown in (S). (S) may look somehow redundant, but it can be useful for people with very low literacy levels [Williams and Reiter 2005].

A	As salas de cinema de todo o mundo exibiam uma produção do diretor Joe Dante em que um cardume de piranhas escapava de um laboratório militar e atacava participantes de um festival aquático. Quase 30 anos depois, (...). Mais de 20 pessoas foram mordidas por palometas (Serrasalmus spilopleura, espécie de piranha) que vivem nas águas da barragem Sanchuri.
B	As salas de cinema de todo o mundo exibiam uma produção do diretor Joe Dante. Na produção, um cardume de piranhas escapava de um laboratório militar e atacava participantes de um festival aquático. Quase 30 anos depois, (...). Mais de 20 pessoas foram mordidas por palometas que vivem nas águas da barragem Sanchuri. Palometas são Serrasalmus spilopleura, espécie de piranha.
C	As salas de cinema de todo o mundo exibiam um filme do diretor Joe Dante. No filme, um cardume de piranhas escapava de um laboratório militar. O cardume de piranhas atacava participantes de um festival aquático. Quase 30 anos depois, (...). Palometas morderam mais de 20 pessoas. As palometas vivem nas águas da barragem Sanchuri. Palometas são Serrasalmus spilopleura, espécie de piranha.

Table 1. An example of an original text (A) and its simplified versions (B and C)

When performing a natural simplification, the annotator is free to choose which operations to use, among the 11 available, and when to use them, although general guidelines suggest shortening of sentences, using canonical order and changing passive into active voice. The annotator can also decide not to simplify a sentence. Strong simplification, on the other hand, is driven by explicit rules from a manual of syntactic simplification also developed in the project [Specia et al. 2008], which explicitly states when and how to apply the simplification operations, with the goal of simplifying the text as much as possible.

The sentence splitting operation, which is the focus in this paper, can be applied usually when a sentence contains apposition, relative clauses, coordinate or subordinate clauses, but it is not a mandatory operation for natural simplifications.

3.2. The parallel corpora of original and simplified texts

The resulting annotated corpora is composed of 104 news articles from *Zero Hora* and 37 from *Caderno da Ciência*. Table 2 shows the total number of sentences in the original, natural and strong simplified versions of the texts. In the simplified version the overall text length is longer than in the original, which was expected, since simplification usually yields the repetition of information in different sentences, particularly when splitting operations are performed.

Table 3 shows the number of sentences with respect to the input texts after the simplifications from original (O) to natural (N), and from natural to strong (S), focusing on the types of operations applied. Most operations can be combined and applied to

Zero Hora			Caderno Ciência		
Original	Natural	Strong	Original	Natural	Strong
2,116	3,104	3,537	569	504	729

Table 2. Number of sentences in the original, natural and strong corpora

the same sentence, except the "Non-simplification" and "Dropping sentence" operations, which are exclusive. In the natural simplification process, the most common syntactic simplification operation is splitting sentences. Strong simplifications (from natural simplifications) prioritize splitting sentences and lexical simplification (lexical substitution and simple rewriting). The high number of non-simplification operations in the strong simplification process is due to the fact that most of the sentences had already been simplified in the natural simplification phase.

Simplification Operations	Zero Hora		Caderno Ciência	
	O → N	N → S	O → N	N → S
Non-simplification	418	2,220	88	231
Strong rewriting	7	4	5	3
Simple rewriting	509	313	113	53
Subject-verb-object ordering	31	13	6	2
Transformation to active voice	89	65	2	43
Inversion of clause ordering	191	74	32	17
Splitting sentences	723	380	59	169
Joining sentences	5	6	4	0
Dropping sentence	6	3	2	1
Dropping sentence parts	241	49	81	24
Lexical Substitution	980	196	322	8

Table 3. Statistics on the simplification operations

4. Natural simplification system

We focus on the sentence splitting operation in this paper, as this is one of the most frequent operations, and can be seen as a key distinctive feature between natural and strong simplification, as shown in Table 3. A binary classifier is trained with a large number of features in order to identify which sentences should be split to produce a natural simplified text, as described in what follows.

4.1. Feature set

From the analysis of our annotated corpora, we extract a number of features which aim to describe the characteristics of the sentences involved (or not) in splitting operations.

Table 4 lists our feature set, which includes superficial, morphological, syntactic and discourse-related features. Features 1 to 26 are considered our *basic* feature set. They reflect findings of previous work and also of our own work within the project, that is, they encode characteristics that are known to influence the complexity of the sentences and consequently its suitability for simplification. Features 2 and 4-18 are similar to the ones proposed by [Petersen and Ostendorf 2007]. The remaining features are based on

#	Feature	#	Feature
1	number of characters	17	average size of VPs
2	number of words	18	number of clauses
3	average size of words	19	number of coordinated clauses
4	number of nouns	20	number of subordinated clauses
5	number of proper names	21	number of relative clauses
6	number of pronouns	22	is there an appositive clause?
7	number of verbs	23	is the sentence in passive voice?
8	number of adjectives	24	number of cue phrases
9	number of adverbs	25	is there a cue phrase in the beginning of the sentence?
10	number of coordinative conjunctions		
11	number of subordinative conjunctions	26	is there a cue phrase in the middle of the sentence?
12	number of noun phrases (NPs)		
13	average size of NPs	27-183	number of occurrences for each cue phrase of a list (157 cue phrases)
14	number of prepositional phrases (PPs)		
15	average size of PPs	184-209	is there a rhetoric relation *x* present in the sentence? (26 rhetoric relations)
16	number of verbal phrases (VPs)		

Table 4. Feature set

lexicalized cue phrases (27 to 183), which include conjunctions and discourse markers such as “assim” and “ao invés de”, and rhetoric relations (184 to 209) (associated with sets of cue phrases) such as “conclusion” and “contrast”. [Williams 2004] has discussed the use of cue phrases in the context of language simplification. The cue phrases and rhetorical relations used here are derived from the ones produced by a discourse analyzer for Brazilian Portuguese [Pardo and Nunes 2006].

Since the cue phrases and rhetorical relations are usually very sparse, we applied different feature selection methods implemented in Weka [Witten and Frank 2005] to keep only the relevant ones: Information Gain, Wrapper, Principal Components, etc. However, these methods did not improve the performance of the classifier. We therefore adopted a simpler feature selection strategy: we trained classifiers using one feature at a time and all features except one at a time (*leave-one-out*), and selected all features that performed above the average accuracy in the first case and which caused a decrease in the classifier’s performance below the average accuracy in the second case. We added the n best performing features selected in this manner to the basic set, experimenting with different values of n on a validation dataset. The best results were obtained with the basic set of features plus the top 50 performing features. The first part of Table 5 lists the 50 selected cue phrases and rhetoric relations. These features do not seem to have been selected based purely on their frequency. For example, the most frequent conjunction, “e”, was not selected and the best performing conjunction, “ou”, is the 15th most frequent. The same applies to the rethoric relations.

4.2. Classification

In order to learn whether to split or not a sentence for natural simplification, we have trained a classifier on the manually annotated corpora. Each sentence in the corpus is represented by the set of features described in section 4.1. Sentences are tagged as positive instances if they were annotated as containing a splitting operation; otherwise they are

ZH	Cues	ou, mas, mesmo, conforme, até, nem, já, com, caso, realmente, principalmente, por isso, para, logo, / (barra), assim, após, que , já que, tanto que, posteriormente, porém, por falar em, em nível de, em geral, em contraste, em contrapartida, em comparação, em adição, é claro, diante de, desse modo, dessa maneira, dessa forma, desde que, de modo semelhante, de fato, daí, da mesma forma, consequentemente
	RR	justify, sequence, antithesis, attribution, comparison, background, summary, circumstance, evidence, conclusion
ZH+CC	Cues	já, porém, como, atualmente, ainda, nem, com
	RR	cause, contrast, list

Table 5. Best performing additional features

tagged as negative.

We use Weka's SMO implementation [Witten and Frank 2005] of Support Vector Machines (SVM) as classification algorithm, with radial basis function kernel and optimized cost and gamma parameters. We have experimented first with the *Zero Hora* corpus, which was the base for our feature selection experiments. It contains 728 examples of the splitting operation and 1328 examples of non-plit sentences, which we randomly split in five different subsamples of training (75%) and test (25%) sets. The training set is further split into validation-train (70%) and validation-test (30%) sets for parameter optimization and feature selection. We report the average performance on these five test subsamples. The first part of Table 6 presents the results of the classification task using four different feature sets for this corpus: (1) the feature set used by [Petersen and Ostendorf 2007], (2) our *basic* set, (3) all our features, and (4) our *basic* set plus the best 50 additional features.

Considering [Petersen and Ostendorf 2007]'s features as our baseline, we show that the features that were added to this baseline yielded a slight increase in the performance of the classifier. The addition of all the discourse-related features contributed to a further small increase in performance. Nevertheless, adding only the top 50 discourse-related features showed considerable improvement with respect to the baseline features. If we compare our results with a simpler baseline, a classifier which always choose the majority class, we observe a large improvement: such classifier obtains Precision of 40.0%, Recall of 63.3% and F-measure of 49.1%.

In a second experiment, we tested our system, trained on the *Zero Hora* corpus, on the *Caderno da Ciência* corpus. We aimed to verify whether our classification strategy is affected but a change in the domain of the texts. The second column of Table 6 shows the results of this experiment. We can observe that the F-measure values are similar to those achieved on the *Zero Hora* corpus; however, there is a considerable increase in precision and a corresponding drop in recall. This may indicate that, besides the simplification patterns learned from the *Zero Hora* corpus, the *Caderno da Ciência* corpus presents additional ones.

In a third experiment, we added instances collected from the *Caderno da Ciência* corpus to our initial training and test data sets (split in the same way as described above). The goal was to verify whether having more training data would improve our performance. The third part of Table 6 presents our results of this experiment. The overall performance using both corpora has increased, proving that more training data can improve

Feature set	Zero Hora			Caderno Ciência			Zero Hora + Caderno Ciência		
	P	R	F	P	R	F	P	R	F
Petersen	71.68	71.54	71.58	81.30	66.70	71.50	76.10	76.48	75.88
Basic	72.48	72.34	72.34	81.20	66.50	71.30	**79.98**	**80.18**	**79.80**
All	72.56	72.48	72.46	80.2	67.5	72.00	77.16	77.44	76.74
Basic+50	**73.50**	**73.42**	**73.40**	80.80	68.00	72.50	77.96	78.24	77.68
Basic+10							**80.52**	**80.72**	**80.26**

Table 6. Results with different feature sets. P=Precision, R=Recall, F=F-measure

the performance of the simplification system. However, using the additional dataset, the best performance was achieved with the basic feature set. This shows that the feature selection process using one corpus (*Zero Hora*) may not be ideal for other corpora.

We decided then to repeat the feature selection procedure as described in Section 4.1 but considering the combined corpus. We again tested adding the best peforming features to our basic feature set, and obtained the best results when adding the first 10 selected features. These results are shown in the last row of the third part of Table 6; these are the best results ofr this dataset. The 10 best performing features according to feature selection using the combined corpus are shown in the second part of Table 5.

4.3. Simplification

The binary classifier described in the previous section only decides whether to split or not a sentence. The actual simplification, when recommended by the classifier, is performed by a rule-based system that implements simplification rules for all syntactic constructions that are considered complex, following the guidelines defined in the Manual for Portuguese Text Simplification [Specia et al. 2008].

The sentence splitting operation is executed when the sentence contains apposition, relative clauses, coordinate or subordinate clauses. However, not all sentences containing these constructs need to be simplified, and therefore the simplification process should rely on a combination of factors. It is the role of our classifier to decide which sentences should be split.

Table 7 shows a few original and (natural) simplified examples where our classifier decided to split or not a sentence. The classifier chose to split sentence (1), but not sentence (2). Both sentences contain relative clauses, but sentence (2) is not a good candidate for splitting because the main clause would become meaningless without the relative clause. We can observe factors that have influenced the correct classification of both sentences (1) and (2): the difference in sentences' lengths, the higher number of clauses and phrases in (1), the longer phrases in (1), the presence/absence of discourse markers. Sentences (3) and (4) are examples of coordinated clauses, but only (3) was chosen to be split. The contributing factors for that decision include the fact that (4) is shorter and has fewer clauses than (3).

5. Concluding remarks and future work

We have presented a corpus-based system for natural text simplification, focusing on the sentence splitting operation as the main point of distinction between this and the strong

1	O	Ele e amigos, como Giovane Silva Ferreira, 13 anos, passam as tardes pescando o peixe, depois levado para uma associação de artesãos **que faz o curtimento da pele do animal**.
	N	Ele e amigos, como Giovane Silva Ferreira, 13 anos, passam as tardes pescando o peixe. Depois, o peixe é levado para uma associação de artesãos. **A associação de artesãos faz o curtimento da pele do animal**.
2	O	Um ser humano, principalmente criança, **que entra em um incêndio sem qualquer treinamento ou proteção**, corre sérios riscos de vida.
	N	Um ser humano **que entra em um incêndio sem qualquer treinamento ou proteção** corre sérios riscos de vida, principalmente se for criança.
3	O	O sol deverá predominar no período, **e as temperaturas mínimas vão variar entre 12ºC e 14ºC em a maior parte de o Estado** – em a Serra, elas continuarão entre 6ºC e 8ºC.
	N	O sol deverá aparecer durante a maior parte das manhãs. **As temperaturas mínimas vão variar entre 12ºC e 14ºC na maior parte do Estado**. As temperaturas mínimas continuarão entre 6ºC e 8ºC na Serra.
4	O	O principal cuidado a ser tomado é usar lente de soldador número 14 **ou projetar com uma luneta a imagem do Sol em uma parede ou em uma cartolina branca**.
	N	No simplification.

Table 7. Split and non-split sentences

level of simplification. We have built a classifier which is able to define when a sentence should or not be split so that the output simplified text is still natural. Our classifier reaches 73.5% precision and 73.4% recall on this task using our best performing feature set when using a data set of general news articles.

We have built a *basic* feature set containing superficial, morphological, and syntactic features, and have experimented adding to these a subset of discourse-related features. This suggests that our discourse-related features contribute to the identification of the sentences that should be split of not for natural simplification.

We have experimented with our simplification classifier on data from another domain. These experiments proved that our approach can handle texts from a different domain, and also that additonal training data can increase the performance of the system.

When our classifier decides for splitting a sentence, the actual splitting and simplification operations are executed by a rule-based system. This simplification framework, corpus-based classifier followed by rule-based simplifier, will be the core of a tool for online simplification of texts on the Web, aiming at people with low literacy levels.

In order to refine the natural simplification classification process, we plan to implement a finer-grained version of the system presented in this paper. Instead of using a classifier to make a decision about the whole sentence (split vs. non split), we aim to have a classification step for each potencial splitting point within the sentence. This would allow us to simplify just specific points of a sentence (assuming that for natural simplification not all syntactic phenomena present in the sentence need to be simplified).

Acknowledgments

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

References

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

Caseli, H. M., Pereira, T. F., Specia, L., Pardo, T. A. S., Gasperin, C., and Aluísio, S. M. (2009). "Building a brazilian portuguese parallel corpus of original and simplified texts". In *Proceedings of CICLing 2009*.

Chandrasekar, R., Doran, C., and Srinivas, B. (1996). "Motivations and methods for text simplification". In *Proceedings of the Sixteenth International Conference on Computational Linguistics (COLING 1996)*, pages 1041–1044.

Chandrasekar, R. and Srinivas, B. (1997). "Automatic induction of rules for text simplification". *Knowledge-Based Systems*, 10(3):183–190.

Devlin, S. and Unthank, G. (2006). "Helping aphasic people process online information". In *Proceedings of the 8th international ACM SIGACCESS Conference on Computers and Accessibility*, pages 225–226, Portland, USA.

Inui, K., Fujita, A., Takahashi, T., Iida, R., and Iwakura, T. (2003). "Text simplification for reading assistance". In *Proceedings of the Second International Workshop on Paraphrasing: Paraphrase Acquisition and Applications*, pages 9–16.

Jr., A. C., Maziero, E., Gasperin, C., Pardo, T. A. S., Specia, L., and Aluisio, S. M. (2009). "Supporting the adaptation of texts for poor literacy readers: a text simplification editor for brazilian portuguese". In *Proceedings of Workshop of Innovative Use of NLP for Building Educational Applications at NAACL 2009*.

Klebanov, B. B., Knight, K., and Marcu, D. (2004). "Text simplification for information-seeking applications". In *On the Move to Meaningful Internet Systems. Lecture Notes in Computer Science*, volume 3290, pages 735–747. Springer-Verlag.

Max, A. (2006). "Writing for language-impaired readers". In *Proceedings of the 7th International Conference on Intelligent Text Processing and Computational Linguistics*, pages 567–570, Mexico City. Springer-Verlag.

Pardo, T. A. S. and Nunes, M. V. (2006). "Review and evaluation of DiZer - an automatic discourse analyzer for brazilian portuguese". In *Proceedings of PROPOR 2006. Lecture Notes in Computer Science*, volume 3960, pages 180–189. Springer-Verlag.

Petersen, S. E. and Ostendorf, M. (2007). "Text simplification for language learners: A corpus analysis". In *Proceedings of the Speech and Language Technology for Education Workshop (SLaTE-2007)*, pages 69–72, Pennsylvania, USA.

Siddharthan, A. (2003). *Syntactic Simplification and Text Cohesion*. PhD thesis, University of Cambridge.

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

Williams, S. (2004). *Natural Language Generation of discourse relations for different reading levels*. PhD thesis, University of Aberdeen.

Williams, S. and Reiter, E. (2005). "Generating readable texts for readers with low basic skills". In *Proceedings of ENLG 2005*, pages 140–147.

Witten, I. H. and Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques, 2nd Edition*. Morgan Kaufmann, San Francisco.