SIMPLIFICA: a tool for authoring simplified texts in Brazilian Portuguese guided by readability assessments

Carolina Scarton, Matheus de Oliveira, Arnaldo Candido Jr.,
Caroline Gasperin and Sandra Maria Aluísio
Department of Computer Sciences, University of São Paulo
Av. Trabalhador São-Carlense, 400. 13560-970 - São Carlos/SP, Brazil
{carolina@grad,matheusol@grad,arnaldoc@,cgasperin@,sandra@}icmc.usp.br

Abstract

SIMPLIFICA is an authoring tool for producing simplified texts in Portuguese. It provides functionalities for lexical and syntactic simplification and for readability assessment. This tool is the first of its kind for Portuguese; it brings innovative aspects for simplification tools in general, since the authoring process is guided by readability assessment based on the levels of literacy of the Brazilian population.

1 Introduction

In order to promote digital inclusion and accessibility for people with low levels of literacy, particularly access to documents available on the web, it is important to provide textual information in a simple and easy way. Indeed, the Web Content Accessibility Guidelines (WCAG) 2.0[1] establishes a set of guidelines that discuss accessibility issues and provide accessibility design solutions. WCAG requirements address not only structure and technological aspects, but also how the content should be made available to users. However, Web developers are not always responsible for content preparation and authoring in a Website. Moreover, in the context of Web 2.0 it becomes extremely difficult to develop completely WCAG conformant Websites, since users without any prior knowledge about the guidelines directly participate on the content authoring process of Web applications.

In Brazil, since 2001, the INAF index (National Indicator of Functional Literacy) has been computed annually to measure the levels of literacy of the Brazilian population. The 2009 report presented a still worrying scenario: 7% of the individuals were classified as illiterate; 21% as literate at the rudimentary level; 47% as literate at the basic level; and only 25% as literate at the advanced level (INAF, 2009). These literacy levels are defined as: (1) **Illiterate**: individuals who cannot perform simple tasks such as reading words and phrases; (2) **Rudimentary**: individuals who can find explicit information in short and familiar texts (such as an advertisement or a short letter); (3) **Basic**: individuals who can read and understand texts of average length, and find information even when it is necessary to make some inference; and (4) **Advanced/Fully**: individuals who can read longer texts, relating their parts, comparing and interpreting information, distinguish fact from opinion, make inferences and synthesize.

We present in this paper the current version of an authoring tool named SIMPLIFICA. It helps authors to create simple texts targeted at poor literate readers. It extends the previous version presented in Candido et al. (2009) with two new modules: lexical simplification and the assessment of the level of complexity of the input texts. The study is part of the PorSimples project[2] (Simplification of Portuguese Text for Digital Inclusion and Accessibility) (Aluisio et al., 2008).

This paper is organized as follows. In Section 2

[1] http://www.w3.org/TR/WCAG20/

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

Proceedings of the NAACL HLT 2010: Demonstration Session, pages 41–44,
Los Angeles, California, June 2010. ©2010 Association for Computational Linguistics

we describe SIMPLIFICA and the underlying technology for lexical and syntactic simplification, and for readability assessment. In Section 3 we summarize the interaction steps that we propose to show in the demonstration session targeting texts for low-literate readers of Portuguese. Section 4 presents final remarks with emphasis on why demonstrating this system is relevant.

2 SIMPLIFICA authoring tool

SIMLIFICA is a web-based WYSIWYG editor, based on TinyMCE web editor[3]. 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 2.2) 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 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.

Figure 1 summarizes how the three modules are integrated and below we describe in more detail the SIMPLIFICA modules.

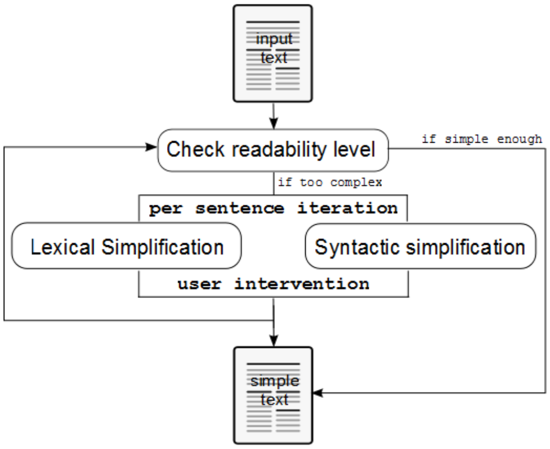

Figure 1. Steps of the authoring process.

2.1 Lexical Simplification

Basically, the first part of the lexical simplification process consists of tokenizing the original text and marking the words that are considered complex. In order to judge a word as complex or not, we use 3 dictionaries created for the PorSimples project: one containing words common to youngsters, a second one composed by frequent words extracted from news texts for children and nationwide newspapers, and a third one containing concrete words.

The lexical simplification module 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].

After the text is tagged, the words that are not proper nouns, prepositions and numerals are selected, and 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 marked 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 listed in order, from the simpler to the more complex ones. 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.

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

[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/

2.2 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 a manual for Brazilian Portuguese syntactic simplification (Specia et al., 2008). 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 had 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 available 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 ones due to sentence splitting. We acknowledge that low-literacy readers prefer short texts, and in the future we aim to provide summarization within SIMPLIFICA (see (Watanabe et al., 2009)). Here, the shortening of the text is a responsibility of the author.

2.3 Readability assessment

With our readability assessment module, we can predict the readability 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[9] toolkit (SMO). We have used 7 corpora within 2 different genres (general news and popular science articles) to train the classifier. Three of these corpora contain original texts published in online newspapers and magazines. The other corpora contain manually simplified versions of most of the original texts. These were simplified by a linguist, specialized in text simplification, according to the two levels of simplification proposed in our project, natural and strong, which result in texts adequate for the basic and rudimentary literacy levels, respectively.

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, 2003)) also developed in the context of the PorSimples project. Coh-Metrix-PORT implements the metrics in Table 1.

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 1. Metrics of Coh-Metrix-PORT.

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

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

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.

We measured the performance of the classifier on identifying the levels of the input texts by a 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. A working session at SIMPLIFICA

In the NAACL demonstration section we aim to present all functionalities of the tool for authoring simple texts, SIMPLIFICA. We will run all steps of the authoring process – readability assessment, lexical simplification and syntactic simplification – in order to demonstrate the use of the tool in producing a text for basic and rudimentary readers of Portuguese, regarding the lexical and the syntactic complexity of an original text.

We outline a script of our demonstration at http://www.nilc.icmc.usp.br/porsimples/demo/demo_script.htm. In order to help the understanding by non-speakers of Portuguese we provide the translations of the example texts shown.

4. Final Remarks

A tool for authoring simple texts in Portuguese is an innovative software, as are all the modules that form the tool. Such tool is extremely important in the construction of texts understandable by the majority of the Brazilian population. SIMPLIFICA's target audience is varied and includes: teachers that use online text for reading practices; publishers; journalists aiming to reach poor literate readers; content providers for distance learning programs; government agencies that aim to communicate to the population as a whole; companies that produce technical manuals and medicine instructions; users of legal language, in order to facilitate the understanding of legal documents by lay people; and experts in language studies and computational linguistics for future research.

Future versions of SIMPLIFICA will also provide natural simplification, where the target sentences for simplifications are chosen by a machine learning classifier (Gasperin et al., 2009).

Acknowledgments

We thank FAPESP and Microsoft Research for supporting the PorSimples project

References

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

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

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.

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 in 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. 2009. Instituto P. Montenegro and Ação Educativa. *INAF Brasil - Indicador de Alfabetismo Funcional - 2009*. Online available at http://www.ibope.com.br/ipm/relatorios/relatorio_inaf_2009.pdf

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.