Computational Approaches for Simplifying Educational Texts: A Proposal Using spaCy

Vitor Amadeu Souza
vitor.souza@ime.eb.br
Military Institute of Engineering
Rio de Janeiro, RJ

ABSTRACT

This work presents an investigation into the application of Natural Language Processing (NLP) techniques for the automatic simplification of educational texts in Brazilian Portuguese. The study uses the spaCy library with the pt_core_news_sm model to perform syntactic analysis, named entity recognition, and textual readability assessment. The proposed methodology implements simplification rules based on syntactic dependency analysis, preserving essential elements such as the subject and main predicate while removing complex subordinate constructions. The results show that named entity analysis was effective in identifying people (PER), locations (LOC), organizations (ORG), and miscellaneous elements (MISC) in the analyzed texts. The original texts presented Flesch Reading Ease scores ranging from 25.23 to 54.57, indicating different levels of complexity. This research contributes to the advancement of automatic text simplification techniques in Portuguese and offers insights for the development of more accessible educational tools.

KEYWORDS

Natural Language Processing; Text Simplification; Syntactic Analysis; Named Entities; Text Readability.

1 INTRODUCTION

Access to quality educational information remains a constant challenge in contemporary society, especially when considering the diversity of literacy levels and textual comprehension among different population groups. According to Scarton and Aluísio [11], automatic text simplification emerges as a fundamental research area for democratizing access to knowledge, allowing complex content to be adapted for different target audiences.

Natural Language Processing (NLP) has become a powerful tool for the automatic analysis and manipulation of texts, offering resources that enable the identification of complex linguistic structures and their subsequent simplification. As highlighted by Jurafsky and Martin [4], NLP techniques make it possible to perform morphosyntactic analysis, named entity recognition, and readability assessment, essential components for text simplification systems.

In the Brazilian educational context, the need for more accessible didactic materials becomes even more evident when considering data from the Functional Literacy Indicator (INAF), which reveals that a significant portion of the population struggles to comprehend complex texts [9]. Therefore, the development of automatic tools for simplifying educational texts represents a relevant contribution to educational inclusion.

This work presents a proposal for the automatic simplification of educational texts using NLP techniques, focusing on syntactic analysis and named entity recognition. The main objective is to develop and evaluate a simplification system that preserves the essential informational content while reducing the syntactic complexity of the analyzed texts.

2 THEORETICAL FRAMEWORK

2.1 Natural Language Processing

Natural Language Processing is an interdisciplinary field that combines computational linguistics, artificial intelligence, and computer science to enable computational systems to understand and manipulate human language. Manning and Schütze [5] define NLP as the set of techniques and algorithms that allow the analysis, understanding, and automatic generation of texts in natural language.

The architecture of NLP systems traditionally involves multiple levels of analysis, including morphological, syntactic, semantic, and pragmatic processing. Bird, Klein, and Loper [2] emphasize that syntactic analysis, in particular, plays a key role in identifying grammatical structures that can be simplified without compromising the original meaning of the text.

2.2 Syntactic Analysis and Dependencies

Dependency parsing represents a fundamental paradigm for the structural understanding of texts. According to Nivre [7], this type of analysis establishes hierarchical relationships between words, identifying elements such as subject (nsubj), direct object (dobj), sentence root (ROOT), and other syntactic dependencies essential for textual comprehension.

The spaCy library, developed by Honnibal and Montani [3], offers efficient implementations of syntactic parsers based on neural networks, enabling accurate identification of dependencies in Brazilian Portuguese texts through the pt_core_news_sm model. This model was trained on corpora specific to the Portuguese language, ensuring greater accuracy in analyzing syntactic constructions typical of Brazilian Portuguese.

2.3 Named Entity Recognition

Named Entity Recognition (NER) is a fundamental NLP task aimed at identifying and classifying textual elements that refer to specific entities such as people, locations, organizations, and other relevant elements. Ratinov and Roth [8] demonstrate that effective NER systems contribute significantly to the semantic understanding of

In: Proceedings of the Brazilian Symposium on Multimedia and the Web (WebMedia'2025). Rio de Janeiro, Brazil. Porto Alegre: Brazilian Computer Society, 2025.

ISSN 2966-2753

texts, facilitating simplification processes that preserve relevant factual information.

In the context of the Portuguese language, Santos and Cardoso [10] highlight specific challenges related to named entity identification, including morphological variations, contextual ambiguities, and cultural specificities that influence the accuracy of recognition systems.

2.4 Automatic Text Simplification

Automatic text simplification is a research area that seeks to reduce linguistic complexity while maintaining the original informational content. Siddharthan [12] categorizes simplification techniques into three main types: lexical simplification (replacing complex words with simpler synonyms), syntactic simplification (restructuring complex grammatical constructions), and semantic simplification (reorganizing information to facilitate understanding).

In the Brazilian context, Aluísio *et al.* [1] were pioneers in developing simplification systems for Portuguese, establishing theoretical and methodological foundations that continue to influence contemporary research in the field. The authors emphasize the importance of preserving essential elements such as the subject and main predicate during syntactic simplification processes.

2.5 Readability Metrics

The assessment of textual readability is a fundamental component for validating the effectiveness of simplification systems. The Flesch Reading Ease index, adapted to Portuguese by Martins *et al.* [6], is a widely used metric that considers factors such as average sentence length and the syllabic complexity of words to determine the reading difficulty level.

Scarton *et al.* [11] demonstrate that readability metrics, when applied in conjunction with deeper linguistic analyses, provide more accurate evaluations of textual accessibility, contributing to the development of more effective simplification systems.

3 METHODOLOGY

3.1 Computational Environment Setup

The development of the simplification system was carried out using Python 3.8, with the following specialized libraries: spaCy 3.4 for natural language processing, WordCloud 1.9.2 for textual frequency visualization, matplotlib 3.5.3 for graph generation, and textstat 0.7.3 for the calculation of readability metrics. The linguistic model pt_core_news_sm was used for the analysis of Brazilian Portuguese texts.

3.2 Analysis Corpus

The experimental corpus consisted of three educational texts representing different knowledge domains: theoretical physics (biography of Albert Einstein and the theory of relativity), space exploration (NASA's Apollo 11 mission), and institutional history (University of Heidelberg). This diverse selection made it possible to assess the robustness of the simplification system across different thematic contexts and syntactic structures. Table 1 presents the analyzed texts.

Table 1: Analyzed Texts

Albert Einstein nasceu em 14 de março de 1879 na Alemanha. Ele revolucionou a física com a teoria da relatividade, que explica como a energia e a massa são equivalentes através da fórmula $E=mc^2$. A relatividade geral, publicada em 1915, reformulou a compreensão da gravidade.
A NASA, fundada em 1958, lançou a missão Apollo 11 em 16 de julho de 1969. A missão levou os astronautas Neil Armstrong e Buzz Aldrin à Lua, marcando o primeiro pouso lunar da história. A agência espacial continua a explorar o universo.
A Universidade de Heidelberg, fundada em 1386, é uma das mais antigas da Europa. Localizada na Alemanha, ela contribuiu para avanços em diversas áreas do conhecimento, como medicina e física.

3.3 Simplification Algorithm

The algorithm processes each sentence individually, identifying syntactic dependencies through spaCy's analysis. It then filters the most relevant tokens based on grammatical criteria such as NOUN (common nouns), VERB (verbs indicating actions or states), and PROPN (proper nouns, such as people, places, and organizations). After this filtering, the system reconstructs the sentences in a simpler form, preserving semantic cohesion and eliminating complex subordinate structures.

3.4 Named Entity Analysis

Named entity extraction was performed using the pt_core_news_sm model, which identifies four main categories: PER (people), LOC (locations), ORG (organizations), and MISC (miscellaneous elements). The quantitative analysis of entities allowed for the evaluation of the informational distribution in the texts and the preservation of factual elements during the simplification process.

3.5 Readability Evaluation

Readability evaluation was conducted using the Flesch Reading Ease index, calculated through the textstat library. This metric produces values between 0 and 100, with higher scores indicating greater ease of reading. The comparison between original and simplified texts allowed for quantifying the impact of the simplification process on textual accessibility.

3.6 Results Visualization

The presentation of results included visualizations of syntactic dependencies and named entities using spaCy's displacy module, as well as the generation of word clouds for the identified entities.

The source code for this experiment is available for download through the following link: https://github.com/vitor-souza-ime/simplification.

4 RESULTS AND DISCUSSION

4.1 Named Entity Analysis

The results of the named entity analysis demonstrated significant effectiveness in identifying relevant factual elements. The system identified 3 people (PER), including Albert Einstein, Neil Armstrong, and Buzz Aldrin; 4 locations (LOC), comprising Germany (twice), the Moon, and Europe; 2 organizations (ORG), specifically NASA and the University of Heidelberg; and 1 miscellaneous element (MISC), corresponding to the Apollo 11 mission.

This distribution reveals the predominance of factual information in the analyzed texts, confirming the educational and informative nature of the selected corpus. The accurate identification of these entities is a key component for simplification processes that preserve essential informational content, as emphasized by Ratinov and Roth [8].

The visualization of named entities through the displacy module showed clarity in automatic identification and categorization, facilitating the understanding of the informational structure of the texts. This capability for automatic recognition represents a significant advantage for large-scale simplification systems, eliminating the need for manual identification of relevant factual elements. Figure 1 presents part of this textual analysis.

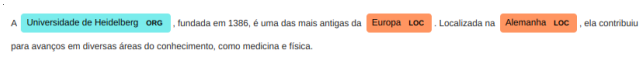

Figure 1: Part of the Named Entities. Source: author.

4.2 Original Readability Evaluation

The readability analysis of the original texts revealed significant variations in the Flesch Reading Ease scores. The text about Albert Einstein presented a score of 41.21, indicating a moderate difficulty level. The text about the Apollo 11 mission obtained a score of 54.57, representing the highest level of accessibility among the analyzed texts. The text about the University of Heidelberg registered a score of 25.23, characterizing it as the most complex in the corpus.

These variations reflect structural and lexical differences among the texts: the Einstein text contains a higher density of technical terms and complex syntactic constructions; the Apollo 11 text maintains a more accessible narrative structure; while the Heidelberg text presents a high informational density in a compact textual structure, resulting in lower readability.

4.3 Syntactic Analysis and Dependencies

The visualization of syntactic dependencies, carried out using the displacy module, revealed complex structures in the original texts, including subordinate constructions, extensive modifiers, and hierarchical syntactic relations. The Einstein text showed greater syntactic complexity, with multiple layers of subordination and modification, while the Apollo 11 text exhibited a more linear and accessible structure.

This analysis confirms the adequacy of the dependency-based approach for identifying complex structural elements, although it highlights the need for refinement in simplification rules to preserve appropriate textual connectivity. Figure 2 shows part of these dependencies.

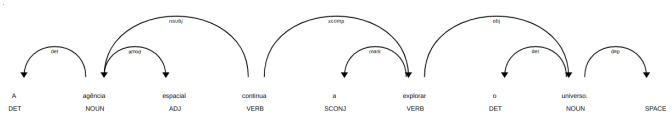

Figure 2: Syntactic Dependencies. Source: author.

4.4 Distribution and Frequency of Entities

The word cloud analysis of the named entities revealed a balanced distribution across different informational categories. Entities such as "Albert Einstein," "NASA," "Germany," and "Apollo 11" displayed visual prominence proportional to their relevance in the analyzed texts, demonstrating the effectiveness of automatic identification of central factual elements.

This ability to automatically identify and quantify entities represents a contribution to simplification systems that need to preserve essential factual information during textual restructuring processes, as illustrated in Figure 3.

Figure 3: Word Cloud. Source: author.

5 CONCLUSION

This study presented an investigation into the application of Natural Language Processing techniques for the automatic simplification of educational texts in Brazilian Portuguese. The results demonstrated effectiveness in the identification of named entities and in the analysis of syntactic dependencies, both fundamental components for text simplification systems.

The named entity analysis revealed a robust capacity for the automatic identification of people, locations, organizations, and miscellaneous elements, preserving essential factual information during text processing. The readability evaluation of the original texts showed significant variations in complexity levels, confirming the need for simplification strategies adapted to different educational contexts.

The contributions of this research include the validation of NLP techniques for analyzing educational texts in Portuguese and the demonstration of the effectiveness of named entity recognition systems for preserving informational content.

Future work may explore hybrid strategies that combine syntactic analysis with more advanced semantic techniques, investigate methods for preserving textual cohesion during automatic simplification, and develop adaptive systems capable of adjusting simplification levels according to the specific characteristics of the target audience.

REFERENCES

[1] Sandra Maria Aluísio, Lucia Specia, Thiago Alexandre Salgueiro Pardo, Erick Gimenes Maziero, and Rodrigo Parreira Maziero Fortes. 2008. Towards Brazilian Portuguese automatic text simplification systems. In *Proceedings of the Eighth ACM Symposium on Document Engineering*. ACM, São Paulo, 240–248.

[2] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Sebastopol.

[3] Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

[4] Daniel Jurafsky and James H. Martin. 2025. *Speech and Language Processing* (3 ed.). Stanford University Press, Stanford.

[5] Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge.

[6] Thereza Bezerra Fraga Martins, Cristina Meneguello Ghiraldelo, Maria das Graças Volpe Nunes, and Osvaldo Novais Oliveira Junior. 1996. *Readability formulas applied to textbooks in Brazilian Portuguese*. ICMC-USP, São Carlos.

[7] Joakim Nivre. 2010. Dependency parsing. *Language and Linguistics Compass* 4, 3 (2010), 138–152.

[8] Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*. Association for Computational Linguistics, Boulder, 147–155.

[9] Vera Masagão Ribeiro. 1997. Alfabetismo funcional: referências conceituais e metodológicas para a pesquisa. *Educação & Sociedade* 18, 60 (Dec. 1997), 143–156. https://doi.org/10.1590/S0101-73301997000300009

[10] Diana Santos and Nuno Cardoso. 2007. *Reconhecimento de entidades mencionadas em português*. Linguateca, Lisboa.

[11] Carolina Eduarda Scarton and Sandra Maria Aluísio. 2010. Análise da inteligibilidade de textos via ferramentas de processamento de língua natural: adaptando as métricas do Coh-Metrix para o português. *Linguamática* 2, 1 (2010), 45–61.

[12] Advaith Siddharthan. 2014. A survey of research on text simplification. *International Journal of Applied Linguistics* 165, 2 (2014), 259–298.