Syntactic Complexity in Students' Essays: A Parsing-Based Analysis*

Renata Ramisch[1,2]

[1] Interinstitutional Center for Computational Linguistics (NILC), São Carlos, Brazil
[2] Graduate Program in Linguistics (PPGL), Federal University of São Carlos, Brazil
renata.ramisch@gmail.com

Abstract. This Master's thesis research proposal intends to investigate the relation between syntactic complexity and writing problems in a corpus of essays by students in the final years of Brazilian regular education. We intend to use Natural Language Processing tools to parse syntactically complex sentences (clauses with at least one embedded sentence within its limits), to investigate the treatment given by these parsers to complex sentences with writing problems, and to analyze the correlation between the complex constructions and automatic parsing results. Then, we will develop a schema of common errors and try to correlate them with the parse trees. We hope these parsing results can be used for the improvement of parsing tools and for the future development of writing aid tools. Thus, we expect to make a significant contribution to the NLP and education fields.

Keywords: Syntactic complexity · Text production · NLP · Parsing.

1 Introduction

Text production has an important status in society and is used in several contexts. Students in the final years of regular education in Brazil should be able to write comprehensible texts, with low average of syntactic problems, to get good grades in university selection exams (e.g. the ENEM - National Exam of High School)[3]. The example bellow shows a sentence written by a student for a mock exam[4]:

> "Segundo o sociólogo Émile Durkheim, é por meio da coerção social que o machismo se perpetua na sociedade, pois uma criança que vive em um meio onde há atitudes preconceituosas *ela tende a adquiri-la*."[5]

* Supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

[3] www.enem.inep.gov.br

[4] Available at https://vestibular.brasilescola.uol.com.br/banco-de-redacoes/13366/

[5] Literal translation: According to the sociologist Émile Durkheim, [no subj] is through social coercion that the sexism perpetuates itself in the society, because a child who lives in an environment where there are prejudiced attitudes *she tends to acquire it*[fem].

The writing problem in the sentence with respect to the Brazilian Portuguese standard language lays in the relative construction "(...) uma criança que vive em um meio [...] *ela tende a adquiri-la*"[6], because it is not possible to identify the reference of the oblique pronoun *-la* in *adquiri-la*. Therefore, it could be interesting to use Natural Language Processing (NLP) to analyze writing problems in students' essays to help both teachers and students to build better texts. For the purposes of the current research proposal, we define *writing problems* as errors according to the Brazilian Portuguese standard language.

The first step towards understanding the students' writing problems could be to study the syntactic complexity of the learners texts. Complex sentences are defined as the ones that contain at least one other embedded sentence within its limits [9]. Our hypothesis is that students might be having trouble writing complex sentences. Therefore, the targeted structures of this analysis are the ones considered syntactically complex (i.e. relative and subordinate clauses). We believe the analysis of such structures could allow us to identify the most relevant difficulties of text producers and, based on these results, make some contribution to the development of writing aid tools for helping them with this task (which is out of the scope of this project).

1.1 Goals

The main goal of the Master's research proposal is to analyze the syntactic writing problems of students' essays based on parse trees generated by automatic parsers, to verify if and how syntactically complex constructions influence the quality of these essays. Further, we intend to verify how automatic parsers deal with sentences that have writing problems, so that these results could help improving these tools and propose guidelines for the development of writing aid tools focused on this type of texts. For this, the work is guided by the following research questions:

- Does the presence of syntactically complex constructions influence the rate of writing problems in a sentence?
- How do automatic parsers treat complex sentences that present writing problems?
- Can some characteristics of the parse trees (e.g. depth, parser scores) indicate the presence of syntactically complex constructions and/or writing problems?
- Can the answers to the questions above be transformed into guidelines for the development of writing aid tools?

2 Related Work

From a readability point of view, text complexity is based on concrete criteria, whether lexical, syntactic or semantic, that make it difficult to understand a

[6] Literal translation: a child who lives in an environment ... *she tends to acquire it*[fem]

given text [4]. Text complexity indicators can be related to readability indexes, lexical frequency values and number of non-literal expressions. Readability and text categorization tools also use metrics that are based on the Fletch index and on other statistical aspects like number and proportion of syllables and words per sentence, and number of simple, passive and subordinate clauses [2,11,3] .

From a text production point of view, there are tools for language learning assessment and assistance that analyze writing errors and propose proofreading and corrections, to improve students writing skills (both for native and non-native speakers) and help teachers in assessing texts [6]. In Applied Linguistics, several works address the question of teaching text production in school [10,14].

3 Materials and Methods

For the validations of the research questions, we intend to divide the study in the following steps:

1. Collection, cleaning and characterization of the corpus: we will compile a set of essays written by students in the final years of regular education in Brazil. One of our possibilities is the compilation of essays from the database of SARESP (Sistema de Avaliação de Rendimento Escolar do Estado de São Paulo)[7], depending on its availability for research purposes.
2. Automatic POS tagging and syntactic parsing: we intend to test at least two different parsers that are available for Portuguese, like UDPipe [13], LX Parser [12], PALAVRAS [1], DepPattern [5] and MaltParser [8] . Since they are very different (e.g. they use dependency or constituency trees), the decision about which tool is the best for our goals will be based on the parsing results. We believe the texts have writing problems that could bring some interesting results for the automatic parsing, that could be used for the improvement of parsing tools in texts with syntactic writing problems.
3. Manual validation and annotation of writing problems and complex constructions: we intend to manually annotate the complex sentences after parsing, verify the treatment given by the parsers and develop a schema of writing problems. For this, we will use annotation platforms such as FLAT [7].
4. Analysis of the parsing results and the correlation between annotations and parse trees: this analysis intends to verify if it is possible to systematize the most common writing problems and correlate them with the results of the automatic parsing. This step will only be possible to define after the preliminary results of the previous steps.
5. Creation of a guide for the development of writing aid tools: based on the research outcomes, we intend to create a guide for the development of writing aid tools focused on syntactic problems.

[7] School Performance Assessment System of the São Paulo State, Brazil: www.educacao.sp.gov.br/saresp/

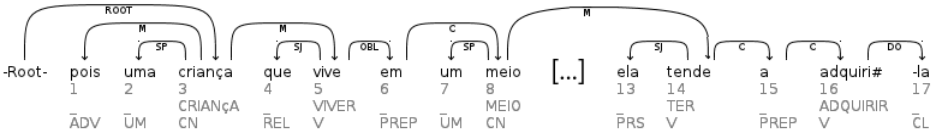
4 Expected Outcomes

This section shows an example of the intended analysis of this research. Given the example shown in Section 1, which presents a complex syntactic constructions (in this case, a relative clause), we used LX DepParser[8] to illustrate the analysis of the relative clause in the sentence. The parsing result[9] is shown as a dependency tree in Figure 1[10].

To compare the parsing treatment, a sentence with the same complex structure, but with no writing problems was also analyzed by LX DepParser. The dependency tree is shown in Figure 2[11]. This example shows the type of problems we expect to find in our corpus.

As we see on the dependency tree, the relative construction in Figure 2 is much closer to a correct syntactic analysis than the one in Figure 1[12]. In the first example, the parser failed in identifying the root (that should be *tende*) and the right syntactic relation between subject (*uma criança* + relative clause) and main verb (*tende*), as indicated by the dependency M (modifier) established between *meio* and *tende*. In the second sentence, the parser was able to successfully identify the root, as well as the syntactic dependency M between the subject (*estudantes*) and the main verb (*têm*) and the secondary dependencies within the relative clause.

Fig. 1. Example of sentence containing relative clause with writing problems.



We have to test other parsers to verify how they analyze such sentences and if they are able to identify the root in Figure 1 correctly. We hope to divide the analysis of the target constructions in three axes:

1. Writing problems: we will verify if the sentence with at least one complex structure **has no writing problems** or if it **has writing problems**.

[8] This tool is not freely available, but it is possible to test specific examples on the web. Available at http://lxcenter.di.fc.ul.pt/services/en/LXServicesParserDep.html

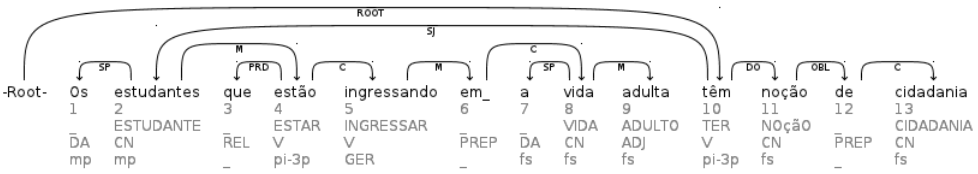
[9] For better readability, a part of the figure has been removed after the parsing process.

[10] Literal translation: because a child who lives in an environment where [...] she tends to acquire it[fem].

[11] Literal translation: The students who are entering in the adult life have notion of citizenship

[12] Both examples are real constructions found on students' essays written for an on-line platform that corrects and evaluates essays for ENEM. Available at https://educacao.uol.com.br/bancoderedacoes/

Fig. 2. Example of sentence containing relative clause without writing problems.



2. Parsing results: we will automatically parse sentences containing the target structures and verify how the parsing tools treat them.
3. Guidelines for writing aid tools: we intend to correlate the two axes above and verify if characteristics of the syntactic trees can provide us with interesting results to be implemented in writing aid tools that deal with syntactic problems.

We expect to be able to quantify the correlation between the target constructions, the presence or not of writing problems, and the treatment given by the parsers of both sentences with and without writing problems. A research possibility outside the scope of this Master's thesis is to use the results to develop a writing aid tool that is not only able to detect the sentences with potential writing problems based on the presence of the target structures and of characteristics of the parse trees, but also to automatically suggest correction alternatives.

5 Conclusions

In this work we propose the study of the relation between syntactic complexity and the writing problems found in students' essays. We believe that, in the long term, these results can be useful for both students and teachers, to use technology to improve the text production and evaluating process.

We also believe that working with large text sets and running automatic analysis with the help of existing NLP tools can speed up text evaluation and bring improvements to the educational context. Technology is present in every single situation of our lives, and computers are able to run some tasks much faster than humans do.

We hope our research can help designing guidelines for the development of writing aid tools and maybe approximate the education and NLP fields with a common goal: helping students to write better texts.

References

1. Bick, E.: The Parsing System PALAVRAS: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Aarhus University Press, Aarhus (Jan 2000)

2. Branco, A., Rodrigues, J., Costa, F., Silva, J., Vaz, R.: Rolling out text categorization for language learning assessment supported by language technology. In: International Conference on Computational Processing of the Portuguese Language, PROPOR. pp. 256–261. Springer (2014)
3. Curto, P.: Classificador de textos para o ensino de português como segunda lıngua. Ph.D. thesis, Master's thesis, Instituto Superior Técnico-Universidade de Lisboa, Lisboa, Lisboa (2014)
4. Finatto, M.J.B.: Complexidade textual em artigos científicos: contribuições para o estudo do texto científico em português. Organon **25**(50) (2011)
5. Gamallo, P.: Dependency Parsing with Compression Rules. In: Proceedings of the 14th International Conference on Parsing Technologies. pp. 107–117. Association for Computational Linguistics, Bilbao, Spain (Jul 2015)
6. Gamallo, P., Garcia, M., del Río, I., González, I.: Avalingua: Natural language processing for automatic error detection. In: Callies, M., Götz, S. (eds.) Learner corpora in language testing and assessment, Studies in Corpus Linguistics, vol. 70, pp. 35–58. John Benjamins Publishing Company, Amsterdam (2015)
7. Gompel, M.v.: FoLiA: Format for Linguistic Annotation. Documentation. Radboud University Nijmegen (2014)
8. Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kübler, S., Marinov, S., Marsi, E.: Maltparser: A language-independent system for data-driven dependency parsing. Natural Language Engineering **13**(2), 95–135 (2007)
9. Perini, M.A.: Gramática descritiva do português brasileiro. Vozes, São Paulo (2017)
10. Santos, P., Motta, V.R.A.: Leitura e produção textual no Ensino Médio: uma proposta a partir da linguística textual. UniLetras **37**(2), 177–186 (2017)
11. Scarton, C.E., Aluísio, S.M.: Análise da Inteligibilidade de textos via ferramentas de Processamento de Língua Natural: adaptando as métricas do Coh-Metrix para o Português. Linguamática **2**(1), 45–61 (Jul 2010)
12. Silva, J., Branco, A., Castro, S., Reis, R.: Out-of-the-Box Robust Parsing of Portuguese. In: International Conference on Computational Processing of the Portuguese Language, PROPOR. pp. 75–85. Springer, Porto Alegre (2010)
13. Straka, M., Straková, J.: Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In: Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. pp. 88–99. Association for Computational Linguistics, Vancouver, Canada (August 2017)
14. Vasconcelos, M.L.M.C., Martins, V.B.: Aulas de produção textual em língua portuguesa no Ensino Médio: um relato de experiência. Revista Teias **18**(49), 144–159 (2017)