UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
CURSO DE CIÊNCIA DA COMPUTAÇÃO

SUYA CASTILHOS

Pylinguistics: an open source library for readability assessment of texts written in Portuguese

Work presented in partial fulfillment
of the requirements for the degree of
Bachelor in Computer Science

Advisor: Prof. Dr. Dante Couto Barone
Coadvisor: MSc. Vinicius Woloszyn

Porto Alegre
July 2016

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
Reitor: Prof. Carlos Alexandre Netto
Vice-Reitor: Prof. Rui Vicente Oppermann
Pró-Reitor de Graduação: Prof. Sérgio Roberto Kieling Franco
Diretor do Instituto de Informática: Prof. Luis da Cunha Lamb
Coordenador do Curso de Ciência de Computação: Prof. Carlos Arthur Lang Lisbôa
Bibliotecária-chefe do Instituto de Informática: Beatriz Regina Bastos Haro

"When you really pay attention, everything is your teacher."

— Ezra Bayda

ACKNOWLEDGMENTS

I would like to thank my university UFRGS and all the professors who helped me learn and inspired me. Even the difficult times were an important part for my personal growth and I will definitely miss my college times.

I would like to thank my friends and family for all the support I received during these years, which definitely helped me keep my mind in place when things got rough.

A personal thanks for my coadvisor Vinicius Woloszyn, for putting up with me and guiding me in this new area I knew so little about. Without him this work wouldn't be possible.

I would also like to thank LaTeX for helping me with the formatting of the text.

ABSTRACT

Readability assessment is an important process in automatic text simplification that aims to measure the text complexity by computing a set of metrics. In this work, I propose the development of a tool with a set of 38 open-source metrics to readability assessment of texts written in Portuguese. I report its development as well as perform a comparative assessment of the tool and obtain satisfactory results with its performance. To illustrate the possibilities of our tool, this work also presents an empirical analysis of readability of Brazilian scientific news dissemination and text categorization.

Keywords: Automatic text simplification. e-accessibility. assessment.

Pylinguistics: uma biblioteca open source para avaliação de inteligibilidade de textos escritos em Português

RESUMO

Avaliação de inteligibilidade é um processo importante na simplificação automática de texto, visando computar um conjunto de métricas capazes de mensurar a complexidade de um texto. Nesse trabalho proponho o desenvolvimento de uma ferramenta de código aberto voltada para a medição de inteligibilidade de textos na língua portuguesa. Reporto o desenvolvimento da mesma, realizo uma avaliação comparativa, e obtenho resultados satisfatórios com sua performance. Para ilustrar as possibilidades da ferramenta, esse trabalho também apresenta uma análise empírica da inteligibilidade do jornalismo científico brasileiro.

Palavras-chave: Simplificação automática de texto. acessibilidade, inteligibilidade.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

POS	Part-of-speech
NLP	Natural Language Processing
UFRGS	Universidade Federal do Rio Grande do Sul
FRE	Flesch reading ease
FGL	Flesch–Kincaid grade leve
ASL	Average sentence length
ASW	Average syllables per word
FSP	Folha de São Paulo

CONTENTS

1 INTRODUCTION

Automatic text simplification is a Natural Language Processing (NLP) task that reduces text syntactic and lexical complexity while preserving, in essence, the original content. The simplified version of the text becomes easier to read and to understand than the original one. This process can be considered a digital inclusion initiative that promotes information access to people with cognitive disabilities (e.g. aphasia and dyslexia) or hearing-impaired people who communicate with each other through the use of sign language. Additionally, text simplification can help people of poor literacy improve their reading skills, including children learning to read different genres of text, second language learners, adults being alphabetized and students undertaking Distance Education, in which text intelligibility is of major importance (HYÖNÄ; OLSON, 1995).

The difficulty of comprehension of a document does not depend only on the linguistic complexity of the text, but also on the reader's reading skills. Some people might easily understand different kinds of text, from scientific papers to intricate novels, while others may find it difficult to read newspaper reports, being necessary an adaptation of the text to their particular personal characteristics. Considering the text aspects, not all texts are of the same genre, and they differ in degrees of complexity. For instance, the Scientific Journalistic genre should exhibit some typical properties, such as relative abstractness, technicality, and informational density while the Journalistic genre destined to general public text should present a higher incidence of nouns and verbs that would decrease the comprehension difficulty of a document (FINATTO et al., 2011).

Related to the problem of automatic simplification is the problem of measuring textual readability with the goal of developing metrics that can associate a readability score to texts. However, to the best of our knowledge, the current readability tools are private or are just available via limited web interfaces that do not allow processing of large amounts of text. This noticeable lack of *public* linguistic tools makes the study of automatic text simplification difficult.

1.1 Goals

The project *Pylinguistics* is the development of a *public*, open source tool for natural language analysis for texts in both English and Portuguese, to serve as a measuring instrument and basis to new studies on the readability area. *Pylinguistics* was proposed and developed by me and MSc. Woloszyn as part of my undergraduate final work at the Computer Science course at Universidade Federal do Rio Grande do Sul (UFRGS).

This work aims to report and document the development and validation of the Portuguese version of *Pylinguistics*, which is already operative. Additionally, to illustrate a use case of the proposed tool, this work also provides an empirical analysis on readability aspects of Brazilian scientific journalism and a comparison with general public journalism, providing an insight on textual characteristics that could make the scientific work more accessible to the general public. Along with other minor tasks, my proposal and main contribution on this work was the developing of features of the tool focused on the Portuguese language and validating it by testing its performance in comparison with a similar, state-of-the-art tool.

1.2 Structure of the text

In Chapter 2, I give some necessary background for the comprehension of the work. In Chapter 3 I describe the field where this research is inserted and present some important related works. Chapter 4 presents the set of metrics chosen and a description of how they were implemented. In Chapter 5, we describe a performance comparison with another similar tool as a mean to validate our tool. Chapter 6 contextualizes a real world problem and illustrates the use of our tool by creating a model for automatically distinguishing journalistic genres. Chapter 7 contemplates possible applications of the tool at its current stage of development and some that would require some small extensions. Lastly, on Chapter 8 I conclude discussing lessons learned in the process as well as a reflection on the contributions and possibilities of further work.

1.3 Complementary files

The files that complement this work, comprising of all the source code and documentation of *Pylinguistics* can be found at <https://github.com/vwoloszyn/pylinguistics>.

2 FOUNDATIONS

In this chapter, I present some background and basic concepts on readability and the technologies involved in the development of this work whose knowledge will help the reader better understand this document.

2.1 Readability

Readability is the name of the ease with which a written text can be understood. The concept of measuring the difficulty of comprehension of a text, as well as measuring how much of the text was successfully understood by the reader is a complex evaluation process, and as defined by Weiss (WEISS; NEUCHÂTEL, 1984), evaluation processes are rather subjective.

Historically, research in this area seeks to measure in objective scores, that which is a quite subjective concept. In this work I do not aim to define or question the definition of readability but rather to explore measures that can help analyze texts in the context of readability, particularly for the Portuguese language.

2.2 Python

The language chosen for the development of *Pylinguistics* was *Python*. *Python* is a high-level general purpose programming language. *Python* supports many programming paradigms, such as object oriented, procedural and functional programming. It's a popular language that is employed in a wide variety of contexts. Particularly in the area of Natural Language Processing, *Python* is a common choice due to many factors:

- **Dynamically typed** The type definition of the data-structures emerge as we code.
- **Comprehensive standard library** Rich built-in support of data-structures.
- **Large amount of open-source code available for use** Many open-source text processing libraries that are easy to integrate and provide many resources for classification, tokenization, stemming, tagging and parsing for both English and Portuguese. Such as Google's Natural Language Toolkit, *NLTK*.
- **High level language** Expressive and succinct, the expressions are intuitive and make the final code easy to read.

- **Scripting language** *Python* code can run in any environment in which there is a *Python* interpreter, making it significantly more portable and easier to use than other programming languages.

All these characteristics make *Python* an ideal choice to easily prototype, test and develop code in the area of *NLP*. Among its few disadvantages there is the fact that it has a slow speed of execution, which can be a problem for some applications, but is a completely manageable issue in the context of this work.

Figure 2.1: Sample code of a *Python* file used for testing *Pylinguistics*

```


import Pylinguistics



text = "Esse processo mostra que estamos diante de um novo modelo de interação unive
das pesquisas é feita nas empresas, mas não dá para a empresa surgir do nada. Mesmo
contratados, a parceria se transformou em uma grande oportunidade de exercício prof
anos."

objpl = Pylinguistics.text(text.decode('utf-8'))

objpl.setLanguage("pt");


print('Features: %s' %objpl.getFeatures())



```

2.3 Natural Language Parsing

A Natural Language Parser is a program which breaks down a string of natural language text into small part-of-speech components relative to the form, function, and syntactic relationship of each part.

For the Portuguese version of *Pylinguistics*, we have decided on *nlpnet* (FONSECA; ROSA, 2013), which is a Python library for Natural Language Processing tasks based on neural networks and specially tailored for working with the Portuguese language. *nlpnet* contains a state-of-the-art parser, performing 97.33% token accuracy on part-of-speech tagging in Portuguese.

2.4 Version control

Version control is a common and useful practice in software development which consists in keeping control of changes on the source code by registering every step of the development process with a tag and timestamp, making easier to visualize, modify and revert modifications.

Version control also provides a documentation of the process of development, with each modification carrying a textual description of the changes being made on the project, making it easier for more than one person to work on the same project. As such, version control is highly recommended and considered a good practice in software development.

In the development of *Pylinguistics* we used *GitHub*, which is a web-based repository hosting service that offers a simple distributed revision control and source code management.

3 RELATED WORK

In this chapter, I briefly present some works on the area of text readability and natural language processing.

3.1 Flesch–Kincaid

The Flesch–Kincaid readability tests are early readability tests that intend to rate how difficult a specific text is to be understood. It consists of two tests, the Flesch reading ease, and the Flesch–Kincaid grade level, both use the same core principle and take only in consideration word length and sentence length. (FLESCH, 1948; KINCAID et al., 1975)

The Flesch-kincaid tests are very limited in the sense that they only analyze two textual features, both being on the lexical level, nevertheless, it remains as some of the most relevant readability tests as of today.

3.1.1 Flesch reading ease

The formula for the Flesch reading ease (FRE) test has had some slight changes over time, the version chosen for *Pylinguistics* is the formula as written by Flesch in the book *How to write plain English* (FLESCH, 1979), which is the currently most used version of the formula. It is as follows

$$FRE = 206.835 - (1.015 \times ASL) - (84.6 \times ASW)$$

Being ASL the average sentence length in words, and ASW the average number of syllables per word.

Values usually (but are not limited to) range between 0 and 100, higher scores indicate that the text is easier to read and lower values mean that the passage is more complex to read. A value between 90 and 100 represents a text suitable for a 5th grade student, while a text that scores below 30 is best understood by college graduates. Table 3.1 shows how the scores translate into US school grades.

Since it's possible to score above 100, some short and simple sentences comprising only one-syllable words usually go above that mark, particularly present in texts aimed

Table 3.1: Flesch scores and the corresponding school grades. Texts with the score shown on the left side will be easy for students on the level shown on the right (FLESCH, 1979).

Score	**School Level**
90 to 100	5th grade
80 to 90	6th grade
70 to 80	7th grade
60 to 70	8th and 9th grade
50 to 60	10th to 12th grade (high school)
30 to 50	college
0 to 30	college graduate

at kids learning to read. The sentence "The cat sat on the mat.", for instance, scores 116. While sentences containing advanced and scientific vocabulary score much lower, being usually best understood by people who study that field. Such as the sentence "The Australian platypus is seemingly a hybrid of a mammal and reptilian creature." which scores 24.4 on the Flesch reading ease formula.

3.1.2 Flesch–Kincaid grade level

The Flesch–Kincaid grade level (FGL) (KINCAID et al., 1975) aims to present a U.S. grade level as an output for the formula, making it easy for teachers and parents to judge the readability of a text or book.

$$FGL = (0.39 \times ASL) + (11.8 \times ASW) - 15.59$$

Being ASL the average sentence length in words, and ASW the average number of syllables per word.

The formula doesn't have an upper bound and can also represent number of years of education required to understand a specific text.

3.1.3 Portuguese adaptation

The Flesch reading ease test has a Portuguese adaptation developed by Martins (MARTINS et al., 1996), which basically consists on adding 42 points on the final result of the original Flesch reading ease formula to compensate for the fact that words in Portuguese typically have a bigger number of syllables and would score as more difficult to read in the traditional Flesch reading ease formula.

This adaptation for Brazilian Portuguese was the metric of choice for *Pylinguistics*' main readability formula for the analysis of texts in the Portuguese language.

3.2 Lexile Framework

The Lexile Framework (LENNON; BURDICK, 2004) is an educational tool that aim not only to rate the text readability level, but the readers reading ability, both using the same scale, named the Lexile Scale. Although it has no upper or lower bounds, it usually ranges from 0-200L (L stands for Lexile), in the case of beginner readers, to above 1700L, for advanced texts (WHITE; CLEMENT, 2001).

The Lexile Framework is not a free software, requiring payment to both analyze a text's readability and the reader's comprehension level. It claims that when the reader's ability and the text are apropriately matches, a reader can enjoy a comprehension rate of about 75%.

Similarly to the Flesch-kincaid tests, it also classifies the text using word length and sentence length as its core principles, but the actual formula used is not disclosed. The formula has been a target for criticism, as many reads that are considered extremely difficult are ranked with a surprisingly low Lexile score, as well as some children's picture books ranking higher than adult novels.

3.3 Coh-Metrix

The works presented before classify the text purely on its lexical features, these properties reveal superficial readability aspects due to their incapacity in capturing other linguistic elements that could aid the reader to connect mentally ideas on the text (DUBAY, 2004). More recently, however, most works focus on different linguistic layers, such as lexical, syntactic, discursive, and conceptual representation to provide a measure of complexity. One example is *Coh-Metrix* (GRAESSER et al., 2004; CROSSLEY; MCNAMARA, 2008), which is a tool for computing the cohesion and coherence of English texts.

The exact definition of coherence and cohesion are subject to much debate. In theory, coherence is defined by the interaction between linguistic representations and knowledge representations in the mind of the reader. While cohesion can be defined as characteristics of the text that are likely to contribute to the coherence of the reader's mental representation (GRAESSER; MCNAMARA; LOUWERSE, 2003)

Coh-Metrix (as of version 3.0) measures 108 indices of the text, including both Flesch measures. The metrics are spread on different linguistic layers, exploring new possibilities for calculating the readability of texts. According to (CROSSLEY; ALLEN; MCNAMARA, 2011), the Coh-Metrix L2 Reading Index, which is a readability score made for classifying texts aimed at second language learners, performs significantly better than traditional readability formulas, which focus only on the lexical aspects of a text.

Coh-Metrix has a free sample version available as a web tool for small texts only, but it is currently not functional due to a server error.

3.3.1 Coh-Metrix-Port

Coh-Metrix-Port (SCARTON; ALUÍSIO, 2010) is a Portuguese adaptation of the original Coh-Metrix ported to Brazilian Portuguese. It implements 48 of the original metrics. The performance of the metrics is intrinsically related to the performance of the POS tagging module used. By default, it uses *PALAVRAS* (BICK, 2000), which is a parser for Portuguese. *PALAVRAS* achieves 99% in terms of morphosyntax (word class and flexion), and 97% in terms of syntax (BICK, 2005).

Coh-Metrix-Port 2.0 is a complete re-write of the original Coh-Metrix-Port coded in python that uses updated resources and tools for improved performance and is available to public use through a web interface, although with limited functionality. This web interface proved very useful for this work, as public tools for natural text processing in Portuguese are hard to find.

3.4 François' AI readability formula

François' "AI readability" formula for French as a foreign language (FRANÇOIS; FAIRON, 2012) is a related work that aims that similarly to this one, aims to adapt all the progress that has been made in readability in English into another language. It doesn't focus on the development of a tool, but rather, on creating a model for classifying readability of texts in the French language. Nevertheless, the paper provides interesting insights on metrics that could prove useful for measuring readability, and was an inspiration for some of the metrics developed in this work.

3.5 Other related works

Other tools are currently being developed in the area of Natural Language Processing, some of which being released simultaneously to this work. I highlight the set of linguistic tools developed by Kyle and Crossley.

These tools comprise a large amount of metrics, focusing on varied aspects of text analysis, such as sentiment analysis, cohesion, psycholinguistic word information and others. However, they are limited to the English language, which indicates the current demand of such tools for the Portuguese language.

- **CLA** CLA is a tool that enables the analysis of texts using very large custom dictionaries. Also, in addition to words, custom dictionaries can include n-grams and wildcards (KYLE; CROSSLEY; KIM, 2015).
- **SEANCE** SEANCE is focused in metrics for sentiment analysis. It includes 254 indices and allows customization of indices, including filtering for particular parts of speech (CROSSLEY; KYLE; MCNAMARA, 2016).
- **TAACO** TAACO is a tool for calculating local and global cohesion. It includes 150 indices, including adjacent overlap indices, and connectives indices (CROSSLEY; KYLE; MCNAMARA, 2015).
- **TAALES** TAALES measures 104 metrics of lexical sophistication, such as indices of frequency, range, academic language, and psycholinguistic word information (KYLE; CROSSLEY, 2015).
- **TAASC** TAASSC is a syntactic analysis tool that measures fine-grained indices of clausal and phrasal complexity, indices of syntactic complexity and frequency-based verb argument construction indices (KYLE, 2016).

4 METRICS

One important issue in text simplification research is the aspects that make a text more or less readable for a target user group. For instance, the PSET project (CARROLL et al., 1998), addressed text simplification for people with aphasia, while the PorSimples project (ALUÍSIO et al., 2008) looked into simplification for people with poor literacy rate. Finally, Finatto et al. (FINATTO et al., 2011) showed the readability aspects of journalistic texts by comparing newspapers geared to two different target audiences.

In *Pylinguistics*, the development of the metrics was guided by previous works on readability aspects of texts (LENNON; BURDICK, 2004; FINATTO et al., 2011; GRAESSER et al., 2004; CROSSLEY; MCNAMARA, 2008; SCARTON; ALUÍSIO, 2010). Many of the early readability studies focus exclusively on lexical features of a text to classify its complexity. In fact, lexical features have been shown to be the most important level of information in readability (CHALL; DALE, 1995; LORGE, 1944). Other levels of information might not lead to predictors of as high efficiency than the lexical level, but combined they can add to an improvement of performance (FRANÇOIS; FAIRON, 2012), as well as providing interesting metrics for further text analysis.

Currently, there are 38 metrics already operative in *Pylinguistics*, divided in six categories: descriptive, word information, diversity, connectives, readability and ratios. In this chapter I present each category with their description, relevance in this work and a list of the metrics comprised. Details of their implementation can be found at the appendices.

4.1 Descriptive

Descriptive metrics provide basic information about the text, such as the number of sentences and words. They help checking the output to make sure that the information makes sense, as well as being important basic values for computing more complex metrics.

- **Word, sentence and syllable count** The total number of words, sentences and syllables in a text.

- **Words per sentence** Some statistical data of the size of sentences in words. We calculate the mean, the median, four percentiles (25, 50, 75 and 90) and also the percentage of sentences above 30 words long.
  The average (mean) sentence length is a classical feature in readability, the work of François (FRANÇOIS; FAIRON, 2012) explored the use of the 90th percentile sentence and inspired the computing of the other percentiles and the median. As well as Daoust (DAOUST; LAROCHE; OUELLET, 1996) inspired the developing of calculating the percentage of sentences above size 30, with the use of such metric in his work for the French language.
- **Syllables per word** Statistical data of the length of words in syllables. We compute the mean, median and four percentiles (25, 50, 75 and 90). Flesch (1948) found that the mean number of syllables per word had a correlation of .66 with comprehension difficulty. In fact, mean word length in syllables combined with mean sentence length are classical readability metrics that make the core principles of the Flesch-Kindcaid and the Lexile readability scores, remaining some of the most important metrics in measuring textual readability.
  The median and percentiles were added by us inspired by the data calculated for sentence size, with the goal of providing extra metrics and possible insights in text analysis.

4.2 Word information

Metrics of word information are based on the concept that to each word is assigned a syntactic part-of-speech category, these categories are separated in content words (adjectives, nouns, verbs and adverbs) and function words (determiners, adpositions[1], pronouns and conjunctions).

Some words can perform multiple syntactic roles. For example, the word “andar” can be a noun (“Seu andar era rápido”) or a verb (“Vou andar hoje”). Through the use of the parser of our choice, *NLPNET* (FONSECA; ALUÍSIO, 2015), which achieves around 97% of accuracy when compared to other state-of-the-art taggers for the Portuguese language (FONSECA; ROSA, 2013). A single part-of-speech category is attributed to each word of the text based on its syntactic context.

[1]adpositions is a cover term for prepositions and postpositions

- **Incidences of part-of-speech elements** We calculate the incidence of word categories (adjectives, nouns, verbs, adverbs, pronouns) per 1000 words in the text, and also the incidence of content and function words per 1000 words. These metrics are reflective of elements of a text that are likely to support a reader's construction of a coherent situation model (GRAESSER et al., 2004).

Figure 4.1: Sample code of *Pylinguistics* comprising of the function that computes the incidence of verbs relative to the total number of words in a text

```python


def verbIncidence(pylinguistObj):
    if (pylinguistObj.postag == []):
        pylinguistObj.postag= tools.getPosTag(pylinguistObj)

    nVerb=0
    for tag in pylinguistObj.postag:
        word = tag[0]
        word_clas = tag[1]
        #if word_clas == "VB" or word_clas == "VBD" or word_clas == "VE
        if word_clas == "VERB":
            nVerb +=1
    #print('adjective %i' %nAdjective)
    verbIncidence=0

    try:
        verbIncidence = nVerb / (float(pylinguistObj.word_count)/1000)
    except:
        verbIncidence = 0

    return verbIncidence


```

4.3 Diversity

Metrics of diversity provide information on the variety of words in the text, a high word diversity means many unique words need to be decoded and integrated with the discourse context, which should make comprehension more difficult. In contrast, if some words are being repeated often in a text, it tends to increase cohesion, and thus, make for a more readable text.

Word diversity is a common metric on the measuring of readability, with works as early as Lorge (1948), who found a correlation between difficulties of passages and the mean frequency of the words in such passages. Both diversity metrics implemented were inspired by Coh-Metrix diversity metrics.

- **Lexical diversity** This metric measures how varied is the total vocabulary of a text. It's defined by the ratio of unique words that appear in the text in comparison to the total number of words in the text.
- **Content diversity** This metric is similar to Lexical diversity, but it only takes in consideration content words (adjectives, nouns, verbs and adverbs). It's defined by the ratio of unique content words that appear in the text in comparison to the total number of content words in the text.

4.4 Connectives

The traditional unit for analyzing grammatical complexity has been the sentence, defined by a staring capital letter and ending on a punctuation mark. However, Coleman (1962) found evidence that independent clauses might be a more valid unit of analysis. Since sentences with connectives such as "Ele acordou e ele foi a aula." (He woke up and he went to school) may actually be treated as if it were two separate syntactic units. The conjunction "e" (and) serves roughly the same function as a punctuation mark. Hence, although the presence of connectives creates longer sentences, their use might decrease the difficulty of understanding of a text by creating cohesive links between ideas and providing clues about text organization (GRAESSER et al., 2004) (Cain Nash, 2011; Crismore, Markkanen, Steffensen, 1993; Longo, 1994; Sanders Noordman, 2000; van de Kopple, 1985).

- **Incidence of connectives** The connectives are divided into five general classes (Halliday Hasan, 1976; Louwerse, 2001), additive, logic, temporal, causal and negative. *Pylinguistics* calculates the incidence of the total number of connectives, as well as the incidence of each separate category. For this, we created a dictionary of connectives in the Portuguese language based on the correspondent connective categories existing in English, as well as some additions based on material found on the web.
- **Logic operators** Within the logical connectives, there are some specific logical particles "e" (and), "ou" (or), "se" (if), and "não", "nem", "nenhum", etc... (not, neither, none, etc...). Measuring such logical particles and how they relate with cohesion and readability in a text was explored by Coh-Metrix-Port (SCARTON; ALUÍSIO, 2010), and inspired 5 metrics of *Pylinguistics*, one

for each group and one metric for the sum of all groups.

Figure 4.2: Sample code of *Pylinguistics* comprising of the function that computes the total incidence of the logical particle “e” (and)

```
def And(pylinguistObj):
    #Incidência do operador lógico E em um texto.
    if (pylinguistObj.language == "pt-br"):
        dic = ['e']
        count =0
        for w in pylinguistObj.tokens:
            w=w.encode('utf-8').lower()
            if w in dic:
                #print(w)
                count+=1
    else:
        dic = ['and']
        count =0
        for w in pylinguistObj.tokens:
            w=w.encode('utf-8').lower()
            if w in dic:
                #print(w)
                count+=1
    #global LogicAnd
    #LogicAnd = count
    #print(count)
    return count
```

4.5 Readability

There are many traditional methods for assessing text difficulty. Klare stated that more than 40 readability formulas have been developed over the years (Klare, 1974-1975). The most common, however, are the Flesch-Kincaid formulas. Since the Flesch reading ease score has a validated Portuguese adaptation (MARTINS et al., 1996), it was our metric of choice for the readability measure in the Portuguese version of *Pylinguistics*.

- **Flesch reading ease, Portuguese version** The Flesch reading ease adaptation for Brazilian Portuguese developed by Martins (MARTINS et al., 1996), basically consists on a shift of 42 points on the result of the original Flesch reading ease formula to compensate for the fact that words in Portuguese typically have a bigger number of syllables.

4.6 Ratios

In contrast to incidence metrics, which compute the number of a specific word type in a span of 1000 words. Ratios are a more relative measure, comparing the incidence of a certain class of units to the incidence of another class of units. Bormuth (1966) demonstrated that the ratios of some part-of-speech elements of the text can be a good predictor of the text's style and genre. At the current state of development, *Pylinguistics* computes a single ratio metric.

- **Ratio of pronouns on prepositions** Inspired on the model of François (FRANÇOIS; FAIRON, 2012) for readability on the French language, we implemented a metric of pronouns on prepositions as a measure of syntactic complexity of sentences. This specific ratio was found to be related to readability of texts in the French language, what signs that it could be an interesting metric to compute for Portuguese.

5 COMPARATIVE ASSESSMENT OF PYLINGUISTICS

In order to provide an assessment of our tool, a comparative evaluation with the similar, state-of-the-art tool *Coh-Metrix-Port* (SCARTON; ALUÍSIO, 2010) was performed. Given that similar tools for the portuguese language are rare and mostly private, I had available only Coh-Metrix-Port 2.0, which is a re-write of the original Coh-Metrix-Port coded in python that claims to use updated resources and tools for improved performance. It is available to public use through a web interface where texts could be individually uploaded and the metrics that result from this text analysis could be downloaded in csv format. However, not all of the features implemented by us in *Pylinguistics* were also available in *Coh-Metrix-Port*, therefore it was just possible to compare a subset of 13 out of our 38 features, those 13 metrics having an exact equivalent metric implemented by *Coh-Metrix-Port*. These 13 metrics are: word count, sentence count, average word per sentence, readability (FRE portuguese adaptation), verb incidence, pronoun incidence, noun incidence, adverb incidence, adjective incidence, connective incidence, content word incidence, functional word incidence and lexical diversity.

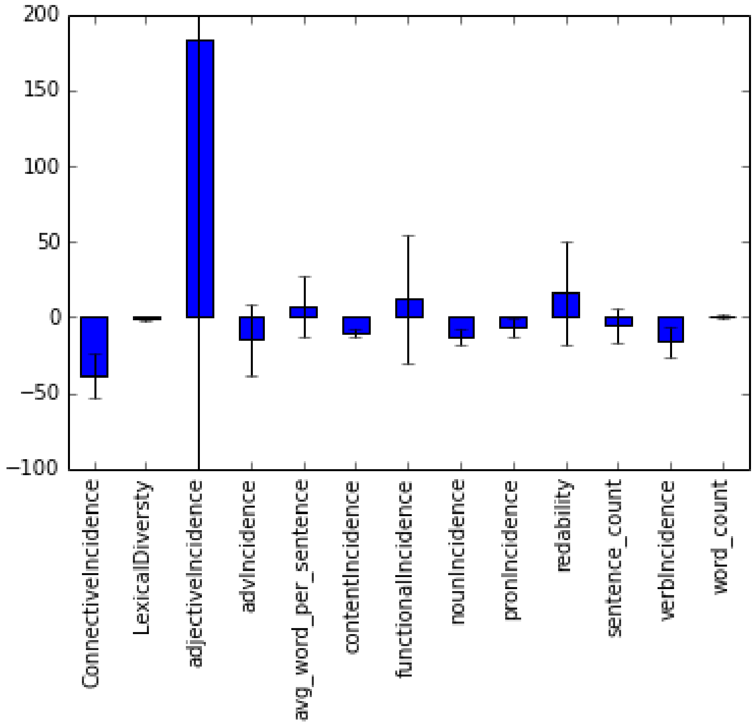
5.1 Methodology

For the comparison, I selected a set of 20 articles from the corpus of Fapesp[1], a magazine in Brazilian Portuguese focused on scientific production. The articles were randomly selected using a random number generator. However, the web interface of *Coh-Metrix-Port* showed some limitations, as it couldn't handle a few articles for being too long. These were substituted by other articles, also randomly selected from the same corpus.

Afterwards, the same articles were processed by *Pylinguistics*, and the results yield were compared with the corresponding results of *Coh-Metrix-Port*. Figure 1 is a box plot representing the difference in results for each metric, with the bar being the average difference in results, positive if greater, negative if smaller than the results of the tool *Coh-Metrix-Port*. The lines represent the standard deviation of the respective values. The results show a similar performance for nearly all metrics, all of them with a difference smaller than 50%, with the exception of the Adjective

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

Figure 5.1: Box plot of the comparison of metrics with *Coh-Metrix-Port*, bars represent mean difference in the result of *Pylinguistics* compared with the results of *Coh-Metrix-Port*, lines represent the respective standard deviation



Incidence metric, in which our tool was classifying a much larger number of words as adjectives than the comparison tool.

5.2 The Adjective disparity

To better understand the disparity in terms of adjective incidence, I performed an exhaustive manual count of all adjectives on the 20 texts of our sample. Afterwards, I calculated the difference from the expected results to the results yield by each tool, as well as the mean and standard deviation of such differences, both shown in Table 1. This analysis lead to the conclusion that *Pylinguistics* adjective labeling comes closer to the number expected from the manual count than *Coh-Metrix-Port*, as well as having a smaller variation.

Since labeling words with their respective part-of-speech tag is a parser task. This difference is most likely due to the difference in terms of the parser used by both tools. For *Pylinguistics* we used the NLP parser, Coh-Metrix-Port 2.0 claims to use *PALAVRAS* (BICK, 2000). Taking in consideration the *PALAVRAS* parser yields correctness rates of over 99% for part-of-speech tagging, I theorize that either Coh-

Metrix-Port 2.0 uses a different parser or there is some bug concerning the adjective count in its code. Especially since other part-of-speech parser tasks such as number of nouns and verbs is well within the expected value.

Table 5.1: Difference in number of adjectives found

Tool	**Mean**	**Standard Deviation**
Pylinguistics	**8.45%**	**8.04**
Coh-Metrix-Port	26.56%	19.59

5.3 Final Remarks

The set of 13 metrics analyzed in this comparative assessment is rather diverse, comprising of some metrics relative to the performance of our parser, some related to counts implemented by us, and also the readability metric, which is the Portuguese adaptation of the Flesch reading ease formula. Since several of the non-tested metrics are relative to the parser's performance, and simple mathematical variations of them (mean, median and percentiles), and the tested metrics showed a satisfactory result in comparison with a state-of-the-art tool, I consider that the results of the comparative assessment were enough for validation of *Pylinguistics*. Other forms of validation are possible, such as performing exhaustive manual counts for the other metrics; buying the license of tools for text analysis in the Portuguese language and performing a comparative assessment again, preferably in comparison with another parser; or validating the English version of equivalent metrics of *Pylinguistics* that don't depend on the parser with the goal of validating the correspondent Portuguese metrics. However, due to time and resource limitations, this forms of validation will not be encompassed in this work.

6 STUDY CASE

To illustrate an application of *Pylinguistics*, we built a model for text categorization. Contextualizing the real world problem of the complexity and intelligibility of the scientific journalism.

Society is large and scientific experiments across the world are carried out by a relatively small number of people, usually hidden from public view in their laboratories. What they do affects everyone, yet most people remain largely unaware of how these scientists use taxpayers' money, and how their work impacts on society. There is, therefore, a considerable need of reporting strategies to communicate scientific advances to the general society. This need largely defines the role of the scientific journalism genre.

However, there are still few computational linguistic studies devoted to observe their textual constitution with particular emphasis on the characterization of stylistics elements of this textual genre. A thorough scientific journalism description can be extremely important for many of the core problems that computational linguists are concerned with. For example, parsing accuracy could be increased by taking genre into account, for instance, certain object-less constructions occur only in recipes in English. Similarly for POS-tagging, where the frequency of uses of *trend* as a verb in the *Journal of Commerce* is 35 times higher than in *Sociological Abstracts*. In information retrieval, genre classification could enable users to sort search results according to their immediate interests, for example scholarly articles about supercollider, novels about the French Revolution, and so forth.

6.1 Methodology

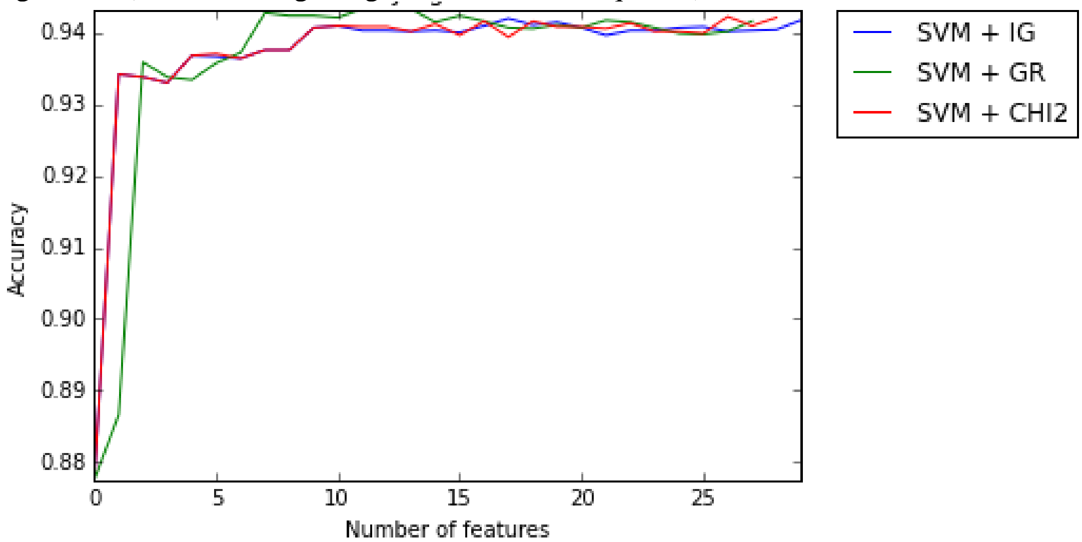
Crossley et al. (CROSSLEY et al., 2007) present a readability analysis using the Coh-Metrix tool by comparing one complex and one simple corpus. Similarly, to illustrate one of the utilities of the proposed tool, we use *Pylinguistics* to compare scientific news dissemination with simple texts to highlight readability features that could make the scientific work more accessible to the general public.

The two corpora used in this study are focused on different groups of readers. Thus, they employ different vocabularies and textual structures that can be classified into different levels of complexity. In this study, we compared two corpora: *Pesquisa*

Table 6.1: Description of the corpora used in this study

Corpus	Articles	Words	Words per article
a) FAPESP	3,866	6,266,831	1,621.01
b) FSP	3,808	1,330,335	349.3

Figure 6.1: Performance of the SVM varying the number of features selected for each algorithm (information gain, gain ratio and chi squared)



Fapesp a Brazilian specialized science magazine; and Folha de São Paulo (FSP) a Brazilian newspaper aimed at the general public. Table 6.1 presents some descriptive statistics on the corpora used in this work.

The FAPESP corpus is composed of articles obtained from the magazine *Pesquisa Fapesp* that has its primary focus on the national (Brazilian) scientific production. It is composed of 3880 articles from 237 editions collected along 19 years. The FSP is a set of 3808 journalistic articles published from 1994 to 1995 by Folha de São Paulo.

6.2 Results

Using the 38 metrics of *Pylinguistics* we extracted the textual features of both corpora, and calculate the average score and standard deviation of each metric for both text categories. With this information in hand, the last step is to select the most informative subset of metrics and combine them into a model for text classification. Several ways of combining the best possible subset of predictors are possible. We decided on using a support vector machine (SVM) since previous works have shown that it yields better results for similar models of text categorization (FRANÇOIS; FAIRON, 2012).

We then assessed the predictive power of our features based on 3 feature selection algorithms commonly used for text categorization: Information Gain, Gain Ratio and Chi-square (DEBOLE; SEBASTIANI, 2004). Figure 2 shows the performance of the SVM when varying the number of features selected for each method. It shows that with only 2 features we can already predict the genre with over 93% accuracy. Additionally by 7 metrics it already reaches the best possible result (over 94%). Finding a small subset of predictors is important to avoid over-fitting, since with a small subset of predictors we have a better chance of generalizing the model to categorize different genres of text.

7 POSSIBLE APPLICATIONS

Even though *Pylinguistics* was conceived with the main concern of measuring textual readability, being a lexically oriented tool for textual analysis it opens a wide possibilities of applications within the area of natural language processing. In this chapter I discuss some of these alternative applications of as well as small additions to *Pylinguistics*' features that could enhance its application for each new goal.

7.1 Content analysis

In principle, *Pylinguistics*' metrics are not oriented to the content of the text, but rather, the form on which the content is presented. However, some lexical and syntactic metrics have been shown to aid insights in content analysis. The book *L'analyse de contenu* (BARDIN, 1977), which is a reference in the area of Content Analysis, presents studies of the content of texts that draw significant conclusions from simple metrics, such as word frequency and verb inflection. Such metrics, like verb inflection, were considered for being part of *Pylinguistics*' set of metrics, and could be developed in the future, being relevant not only for possible applications in the area of content analysis, but also for the measuring of readability (FRANÇOIS; FAIRON, 2012).

The addition of dictionaries of categories is also a likely update in *Pylinguistics*' metrics. Computing incidences of words in pre-defined categories is historically one if the main process of content analysis (BARDIN, 1977).

7.2 Text categorization

A comprehensive set of metrics can be used to better understand trends in different literary genres, as well as automatizing the process of categorizing text. As such, *Pylinguistics* provides a useful tool for the creation of models that could extract the necessary information from the text, even when readability is not the main concern. Text categorization, particularly, can be used to enhance further analysis, as parser accuracy can be increased by taking genre into account.

In fact, *Pylinguistics* has already been used in a study conducted in parallel with this work about categorizing different journalistic genres with the goal of providing insights between the difference in style of scientific and popular journalism (CASTILHOS; WOLOSZYN; BARONE, forthcoming), a disparity in writing form that makes the scientific journalism of difficult comprehension for a large part of society. The model developed in such study is presented in section 6.

7.3 Sentiment analysis and Behavior prediction

In theory, *Pylinguistics* could be used to enhance the results of models for sentiment analysis and behavior prediction, typically the area of sentiment analysis involves a lexicon-based approach whose effectiveness strongly depends on the performance of the lexical resource it relies on(MUSTO; SEMERARO; POLIGNANO, 2014). Such lexicon-approach can benefit *Pylinguistics*, since it provides an interesting perspective and relevant metrics that can possibly aid the measuring of a text's readability. Therefore, the developing of such update can not only boost *Pylinguistics*' main goal of readability, but make it a valid application for new areas of textual analysis.

8 CONCLUSION AND FUTURE WORK

The tool *Pylinguistics* was introduced in this work. It is a public, open source tool for natural language analysis for texts in both English and Portuguese. It is able to extract and compute several metrics from texts.

The goal of *Pylinguistics* is to provide a wide range of textual measures for anyone interested in text analysis, with special focus on the readability aspects of texts. Large amounts of texts can be evaluated by the tool in a short time, and since it was conceived as a public *Python* library, it is of easy integration with other applications. The tool can be used as part of a textual simplification process, functioning as a digital inclusion initiative to help people with cognitive disabilities, people who are learning to read a different genre of text and language learners in general enjoy a more comprehensive understanding. But also as a stand-alone tool to measure readability, essentially providing an accessible way for teachers, students, researchers and parents who feel the need to quantify a text's difficulty of comprehension.

Besides, since it is an open source project it can also be easily extended to suit varied needs, such as content analysis, text categorization, sentiment analysis and behavior prediction, as addressed in 7.

This work aimed to describe, document and validate the developing of *Pylinguistics* Portuguese version of the metrics (Brazilian Portuguese, specifically). Which is specially relevant considering the scarcity of available tools for this language.

To verify the functionality of the tool for the Portuguese language I performed an comparative assessment with a similar, state-of-the-art tool. This way, it was possible to conclude that the Portuguese metrics of *Pylinguistics* show similar results in terms of values computed than the baseline tool, and performing better than the latter in one particular metric that was tested against a specialized manual analysis.

Previous works already have reported the use of linguistic metrics to provide comparison and understanding of the adequacy of text to a target audience (FLESCH, 1948; MARTINS et al., 1996; GRAESSER et al., 2004; CROSSLEY; MCNAMARA, 2008; SCARTON; ALUÍSIO, 2010). However, existing tools are either private, of limited use or simply not available at all in the Portuguese language (GRAESSER et al., 2004; CROSSLEY; MCNAMARA, 2008; LENNON; BURDICK, 2004; SCARTON; ALUÍSIO, 2010). The great number of current alternatives that tackle the issue of measuring textual readability, most of which being commercial products, demonstrate the relevance of this work in the field.

Thus, the positive results of *Pylinguistics*' performance in comparison to an existing tool not only justify this work but open the possibility of future improvement of the tool. I particularly highlight the future development of a few additional metrics, such as a dictionary of simple words, new ratios of part-of-speech elements and the L2 index for foreign language students, all of which are currently in process of development.

There is also the intention of developing an application program interface (API) to provide access to the tool via web, so that the users can input text directly through their browser and see the results in real time. This simple improvement will make the tool easily available to the general public, which not only opens several new possibilities of use but will also definitely help promoting it, making it reach a broader public of possible users.

REFERENCES

ALUÍSIO, S. M. et al. Towards Brazilian Portuguese automatic text simplification systems. In: **ACM symposium on Document engineering**. [S.l.: s.n.], 2008. p. 240–248.

BARDIN, L. **L'analyse de contenu**. Presses universitaires de France, 1977. (Collection 'Le Psychologue'). Disponível em: <https://books.google.com.br/books?id=afcD3h0WL0kC>.

BICK, E. **The parsing system" Palavras": Automatic grammatical analysis of Portuguese in a constraint grammar framework**. [S.l.]: Aarhus Universitetsforlag, 2000.

BICK, E. Gramática constritiva na análise automática de sintaxe portuguesa. **A Língua Portuguesa no Computador. Campinas: Mercado de Letras, São Paulo: FAPESP. ISBN**, p. 85–7591, 2005.

CARROLL, J. et al. Practical simplification of english newspaper text to assist aphasic readers. In: **AAAI Workshop on Integrating Artif Intel and Assistive Tech**. [S.l.: s.n.], 1998. p. 7–10.

CASTILHOS, S.; WOLOSZYN, V.; BARONE, D. Pylinguistics: an open source library forreadability assessment of texts written inportuguese. **iSys - Revista Brasileira de Sistemas de Informação**, forthcoming.

CHALL, J. S.; DALE, E. **Readability revisited: The new Dale-Chall readability formula**. [S.l.]: Brookline Books, 1995.

CROSSLEY, S. A.; ALLEN, D. B.; MCNAMARA, D. S. Text readability and intuitive simplification: A comparison of readability formulas. **Reading in a foreign language**, University of Hawaii, National Foreign Language Resource Center, v. 23, n. 1, p. 86, 2011.

CROSSLEY, S. A.; KYLE, K.; MCNAMARA, D. S. The tool for the automatic analysis of text cohesion (taaco): Automatic assessment of local, global, and text cohesion. **Behavior research methods**, Springer, p. 1–11, 2015.

CROSSLEY, S. A.; KYLE, K.; MCNAMARA, D. S. Sentiment analysis and social cognition engine (seance): An automatic tool for sentiment, social cognition, and social-order analysis. **Behavior research methods**, Springer, p. 1–19, 2016.

CROSSLEY, S. A. et al. A linguistic analysis of simplified and authentic texts. **The Modern Language Journal**, Wiley Online Library, v. 91, n. 1, p. 15–30, 2007.

CROSSLEY, S. A.; MCNAMARA, D. S. Assessing l2 reading texts at the intermediate level: An approximate replication of crossley, louwerse, mccarthy & mcnamara (2007). **Language Teaching**, Cambridge Univ Press, v. 41, n. 03, p. 409–429, 2008.

DAOUST, F.; LAROCHE, L.; OUELLET, L. Sato-calibrage: Présentation d'un outil d'assistance au choix et à la rédaction de textes pour l'enseignement. **Revue québécoise de linguistique**, Université du Québec à Montréal, v. 25, n. 1, p. 205–234, 1996.

DEBOLE, F.; SEBASTIANI, F. Supervised term weighting for automated text categorization. In: **Text mining and its applications**. [S.l.]: Springer, 2004. p. 81–97.

DUBAY, W. H. **The Principles of Readability A brief introduction to readability research**. [S.l.]: Impact Information, Costa Mesa, CA, 2004.

FINATTO, M. J. B. et al. Características do jornalismo popular: avaliação da inteligibilidade e auxílio à descrição do gênero. In: **Brazilian Symposium in Information and Human Language Technology**. [S.l.: s.n.], 2011.

FLESCH, R. A new readability yardstick. **Journal of applied psychology**, American Psychological Association, v. 32, n. 3, p. 221, 1948.

FLESCH, R. **How to write plain English: a book for lawyers and consumers**. Harper & Row, 1979. ISBN 9780060112783. Disponível em: <https://books.google.com.br/books?id=-kpZAAAAMAAJ>.

FONSECA, E. R.; ALUÍSIO, S. M. A deep architecture for non-projective dependency parsing. In: **Proceedings of NAACL-HLT**. [S.l.: s.n.], 2015. p. 56–61.

FONSECA, E. R.; ROSA, J. L. G. Mac-morpho revisited: Towards robust part-of-speech tagging. In: **Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology**. [S.l.: s.n.], 2013. p. 98–107.

FRANÇOIS, T.; FAIRON, C. An ai readability formula for french as a foreign language. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning**. [S.l.], 2012. p. 466–477.

GRAESSER, A. C.; MCNAMARA, D. S.; LOUWERSE, M. M. What do readers need to learn in order to process coherence relations in narrative and expository text. **Rethinking reading comprehension**, p. 82–98, 2003.

GRAESSER, A. C. et al. Coh-metrix: Analysis of text on cohesion and language. **Behavior research methods, instruments, & computers**, Springer, v. 36, n. 2, p. 193–202, 2004.

HYÖNÄ, J.; OLSON, R. K. Eye fixation patterns among dyslexic and normal readers: effects of word length and word frequency. **Journal of Experimental Psychology: Learning, Memory, and Cognition**, American Psychological Association, v. 21, n. 6, p. 1430, 1995.

KINCAID, J. P. et al. **Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel**. [S.l.], 1975.

KYLE, K. Measuring syntactic development in l2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication. 2016.

KYLE, K.; CROSSLEY, S. A. Automatically assessing lexical sophistication: Indices, tools, findings, and application. **TESOL Quarterly**, Wiley Online Library, v. 49, n. 4, p. 757–786, 2015.

KYLE, K.; CROSSLEY, S. A.; KIM, Y. J. Native language identification and writing proficiency. **International Journal of Learner Corpus Research**, John Benjamins Publishing Company, v. 1, n. 2, p. 187–209, 2015.

LENNON, C.; BURDICK, H. The lexile framework as an approach for reading measurement and success. **electronic publication on www. lexile. com**, 2004.

LORGE, I. Predicting readability. **The Teachers College Record**, Teachers College Record, v. 45, n. 6, p. 404–419, 1944.

MARTINS, T. B. et al. **Readability formulas applied to textbooks in Brazilian Portuguese**. [S.l.]: Icmsc-Usp, 1996.

MUSTO, C.; SEMERARO, G.; POLIGNANO, M. A comparison of lexicon-based approaches for sentiment analysis of microblog posts. **Information Filtering and Retrieval**, p. 59, 2014.

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**, v. 2, n. 1, p. 45–61, 2010.

WEISS, J.; NEUCHÂTEL. **La subjectivité blanchie?** Institut romand de recherches et de documentation pédagogiques, 1984. (IRDP/R). Disponível em: <https://books.google.com.br/books?id=BOqsHAAACAAJ>.

WHITE, S.; CLEMENT, J. **Assessing the Lexile Framework: Results of a panel meeting**. [S.l.]: US Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics, 2001.