Letras

Revista do Programa de Pós-Graduação em Letras
Universidade Federal de Santa Maria

O Uso de Metodologias de Corpus em Pesquisa de Linguística Aplicada n.62

v. 31, n. 62, jan./jun., p. 05-21, 2021 - DOI:10.5902/2176148568033 - ISSN 2176-1485

Artigo Original

Submissão: 09/10/2021 Aprovação: 15/04/2022 Publicação: 19/08/2022

A comparison of syntactic patterns of Brazilian Portuguese and English: using comparable corpora of museum texts

Uma comparação dos padrões sintáticos do português brasileiro e do inglês: um exemplo com corpora comparáveis de textos de museus

Lucas Tcacenco

Universidade Federal do Rio Grande do Sul

Sabrina Bonqueves Fadanelli

Universidade de Caxias do Sul

Abstract: The aim of this paper is to analyze two corpora (one in Brazilian Portuguese and another in English), from PUCRS' Science and Technology Museum, to unveil which of these languages could be considered more or less syntactically and rhetorically elaborate. The corpora were processed on Natural Language Processing and Corpus Linguistics tools, such as NILC-Metrix, Coh-Metrix 3.0 and AntConc. Results point to the feature regarding the higher number of words before the main verb as the main factor for syntactic and rhetoric elaboration and, consequently, more complexity in the Portuguese corpus.

Keywords: Corpus linguistics; Syntactic features; Textual complexity; Natural language processing; Museum texts

Resumo: O propósito deste trabalho é analisar dois corpora (um em português brasileiro e outro em inglês), oriundos do Museu de Ciências e Tecnologia da PUCRS no intuito de evidenciar qual dessas línguas poderia ser considerada mais elaborada sintática e retoricamente. Utilizaram-se ferramentas de Processamento de Linguagem Natural e Linguística de Corpus, tais como NILC-Metrix, Coh-Metrix 3.0 e AntConc, para processar os corpora. Os resultados indicam que o número de palavras antes do verbo principal desponta como fator de maior elaboração sintática e retórica, atribuindo maior complexidade ao corpus em português.

Keywords: Linguística de Corpus; Sintaxe; Complexidade textual; Processamento de linguagem natural; Textos de museus

Introduction

It is an undisputed fact that the advances in science and technology have a major impact on our lives. For instance, today we can communicate with people from around the world just by a single click on a smartphone. Similarly, when we look at the efforts scientists make in devising methods to divert away potential near-Earth objects from colliding with planet Earth, or even at the feeling of relief when a vaccine for a deadly disease is produced, we can realize that without science and technology, our existence would likely be in danger.

However, some scientific advances are not that easily available to a large number of individuals, and the reasons for that are various. By way of illustration, a lot of people do not have access to the channels through which science is popularized; to make matters worse, the language presented in science texts may not be suiChart to the language needs of certain audiences.

There are a number of institutions that somehow devote themselves to presenting the advances of science and technology, such as science centers and science museums. These institutions' exhibits tend to complement the instruction given at school, or which may be presented in other channels. One such type of museum is PUCRS' Science and Technology Museum (MCT-PUCRS), in Porto Alegre, Brazil. That museum is a reference for the popularization of science and technology in that country, as it welcomes hundreds of people, from the world over, every day, to learn more about science and technology in a fun and interactive way. As a consequence, most of MCT-PUCRS' exhibits are presented in two languages, Brazilian Portuguese and English, to cater to wider audiences.

One of the key factors to make a visiting experience at a museum fun and interactive is the ability to present the language in a way that is easy to understand for its intended audience. Otherwise, the rich experience of visiting a museum would run the serious risk of being a burden, especially for lower-literacy individuals. Then, great efforts have to be made on the part of museologists and museum curators alike to cater to those audiences.

This entails, among other things, presenting language in an accessible format, especially to individuals whose literacy is under development, which is the Brazilian reality. According to Tcacenco, Silva and Finatto (2019) the INAF numbers (an index for functional illiteracy) state that only 12% of the literate Brazilian population is able to understand complex texts. The same authors bring a discussion on how lexical features that should be aimed at a text's target audience may be presented in a complex text, making comprehension more difficult. The important issue of textual and terminological accessibility in different technical-scientific domains has been covered by other recent works of research, aiming at aiding the improvement of accessible text production and the translation of these texts into a foreign language (RODRIGUES; FREITAS; QUENTAL, 2013; FINATTO; MOTTA, 2019; CORTINA SILVA; DELGADO; FINATTO, 2021)

In an effort to unveil the factors that could potentially make a corpus of texts written in a given language more or less rhetorically and syntactically elaborate than a corpus in another language, two corpora of 50 MCT-PUCRS texts each have been collected. The first corpus is in Brazilian Portuguese (Portuguese corpus) and consists of 50 texts presented at MCT-PUCRS, accompanying its

exhibits. The second corpus consists of the English versions of the texts in the Portuguese corpus (hence English corpus). Our assumption is that the English versions are less elaborate in terms of syntax and discourse and, as a result, less complex. We also aim at checking if some specific syntax features of the corpus may explain why this difference in complexity might be present in the original Portuguese texts and the translated versions in English.

On the lines that follow, an overview of Corpus Linguistics is given. Then, the sub-field of Natural Language Processing (NLP) is explored in more detail. A brief overview of MCT-PUCRS and its texts is presented. After that, the methodology and the tools used in the study are outlined, and then the reader will find a discussion about the results.

Corpus Linguistics

John Sinclair (1991) and the Cobuild project at the University of Birmingham (UK) are the ones known as the starting gear for Corpus Linguistics as it is carried out today since modern technology with computerized analysis has been established. Corpus Linguistics is based on empirical analysis, on existing patterns in natural texts, using software tools to observe the behavior of the language in its natural habitat – the text (BERBER-SARDINHA, 2004; MCENERY, 2012). According to Gries (2010), Corpus Linguistics allows for more truly grounded research patterns in the language under analysis.

Corpus Linguistics, among several other applications, has been also applied to Translations Studies as a reliable approach or methodology to investigate issues that concern the translation process (TAGNIN, 2015).

A more conservative approach to Corpus methods indicates that they are largely associated with massive amounts of texts. There are, however, other approaches which state that a corpus not necessarily has to be giant to be representative (BIBER, 1993). Recent studies (FINATTO, 2018) point to this direction. Findings of her investigation on a corpus of texts on occupational lung diseases indicate that a carefully selected representative small corpus can provide relevant information about the language under study and the genre.

For this study, AntConc (ANTHONY, 2005) is the Corpus Linguistics tool that will be used. It is a free concordance tool to study large corpora that can provide word lists of a given corpus, as well as information about the keywords, collocates and n-grams. The grammatical categories under analysis will be adverbs and verbs. Tagged versions of both corpora will be processed in an effort to facilitate the analysis of these language features. The taggers used were CLAWS part-of-speech tagger[1] for the texts in English, and the LX Parser[2] for the texts in Portuguese.

The following section presents Natural Language Processing (NLP), a sub-field of Artificial Intelligence, and some other tools used in this study.

1 Available at http://ucrel.lancs.ac.uk/claws/. Access in September 2021
2 Available at https://portulanclarin.net/workbench/lx-suite/ . Access in September 2021.

Natural Language Processing – NLP

Corpus Linguistics is able to provide researchers with valuable insights to analyze language phenomena. However, it has a number of limitations (WEIGAND, 2004; ANTHONY, 2013). In an effort to transcend them, a variety of other resources are available to researchers and practitioners alike, and these include the applications of Natural Language Processing – a subfield of Artificial Intelligence, concerned, among other things, with mediating human-computer interaction through language and devising formal methods to analyze texts and produce sentences as if they were uttered by natural people. The efforts of NLP practitioners are channeled to a variety of purposes, which include context recognition, syntactic and semantic analysis, information extraction, machine translation, creation of abstracts, among other things (EVERS, 2013).

NLP is an essentially applied area of interdisciplinary research. Linguistics, on the other hand, navigates from theoretical to applied and vice-versa. Because of their different natures and epistemologies, it is not uncommon for the practitioners of these fields to come into conflict with one another. One general claim on the part of NLP practitioners is that linguists tend to focus too little on the resolution of problems. Linguists, on the other hand, tend to see the linguistic thinking of NLP practitioners a bit too "naïve" (FINATTO; LOPES; CIULLA, 2015). However, apart from these "conflicts", the efforts of these two disciplines have proven to be of great benefit to analyze language.

Among the variety of applications developed in this area, one that deserves distinction in this study is Coh-Metrix (GRAESSER *et al* 2004), a computational tool that produces indices of linguistic and discourse representations of texts through a number of metrics. When analyzed correctly, these metrics can provide information about these texts in terms of lexicon, syntax, semantics, referential cohesion, among other aspects, and can give an indication of how complex they may be.

Coh-Metrix has been adapted for Brazilian Portuguese a number of times, the most recent version being NILC-Metrix (LEAL *et al*). This version, developed by NILC-USP (Inter-institutional Center for Computational Linguistics) of the Universidade de São Paulo, consists of 200 metrics that will similarly assess cohesion, coherence and complexity of texts written in Brazilian Portuguese.

In the following section, the items under analysis - texts presented in science and technology museums – are presented.

The Items under Analysis: Texts Presented in Science and Technology Museums

As mentioned in the Introduction, science and technology museums, such as MCT-PUCRS, do a great service for the popularization of science. The experiments that are available in these institutions can, perhaps, complement the knowledge students gain in school. In fact, the majority of visitors to that museum are Elementary, Middle and High School students.

The experiments visitors interact with at MCT-PUCRS are usually accompanied by a written material, which could be a set of instructions or even information about a given phenomenon. Other

texts, however, are presented in panels with no experiments for visitors to interact with. Today, MCT-PUCRS has a variety of exhibits in different areas, such as Biology, Physics, Archaeology, to name a few.

Below is a sample of a text, originally written in Portuguese, that accompanies an experiment at MCT-PUCRS' bilingual exhibits:

Chart 1 – Sample of an MCT-PUCRS text in Portuguese

ESPELHO SEMITRANSPARENTE
Posicione-se deste lado da lâmina de vidro e peça para um amigo posicionar-se no lado oposto.
Toque o botão (+) para aumentar, gradativamente, a luz sobre seu colega e diminuir a que incide sobre você.
Toque o botão (-) para fazer o contrário.
Percebeu alguma diferença?
Você sempre enxerga o que está mais iluminado. Logo, com mais luminosidade do seu lado, o vidro se comporta como um espelho.

Source: MCT-PUCRS

On the other hand, below is a sample of an English version of a text that accompanies an experiment at MCT-PUCRS' bilingual exhibits:

Chart 2 – Sample of an MCT-PUCRS text in English

HUMAN GYROSCOPE
Because the Soviets arrived first in the space race, Americans decided to work on a project called Project Mercury, from 1958 to 1963. Alan Shepard, on spacecraft Mercury-Redstone 3, was the first American to travel into space, in May 1961. This was only a month after Yuri Gagarin's first flight.
Spacecraft Mercury-Redstone 3 had the shape of a cone and was just large enough for only one person. The mission flew 486 Km in a little more than 15 minutes, at a speed of 8,340 km/h.
The United Stated did not beat the Soviet Union in the Space Race, but Project Mercury was important for men to come to the moon in the near future.

Source: MCT-PUCRS

The idea behind MCT-PUCRS' exhibits is to introduce the lay audience to the advances of science and technology in a language that is as simple as it can be (PUCRS, 2019). On that line, a good starting point for the production of texts that can be potentially non-complex is the principle of Textual and Terminological Accessibility (ATT). According to Finatto *et al* (2016), this principle is a desired property of a text, which would make it understood by readers with some kind of limitation, including their literacy level. This principle has been employed in a number of studies in Brazil (CORTINA SILVA; DELGADO; FINATTO, 2021; FINATTO; TCACENCO, 2020). However, neither of these studies looked at the language patterns used in interlingual[3] translation.

In what follows, we will present the materials and methods employed in this study.

3 Interlingual translation is here understood as the translation in which the source language is different from the target language (JAKOBSON, 1959; ZETHSEN, 2009)

Materials and Methods

Two corpora of texts on exhibit at PUCRS' Science and Technology Museum have been collected for the purposes of this study: 50 of these texts are in English, whereas the other 50 are in Portuguese. Both versions accompany the experiments or are on display in panels. The 100 texts present information on different areas of relevance to science and technology, including Physics, Geography, Sports, History, among others.

There are currently about 600 science and technology experiments and panels at MCT-PUCRS. This means that there are about 600 original texts on display. Although most of the experiments are presented in two languages – Portuguese and English – some of them are presented solely in Portuguese. As our focus was on the comparative study of sentence patterns in the aforementioned languages, 50 texts in Portuguese, and their 50 translations into English, have been carefully selected, for a period of 2 weeks, in August 2021.

The tool NILC-Metrix was used to analyze the information on the 50 texts in Portuguese. NILC-Metrix's English counterpart, Coh-Metrix 3.0, was used to analyze the information on the 50 other texts in English. The metrics that have been selected for this study, however, look at the same features of language (readability, lexicon and syntax, respectively). They are as follows:

a) **Type-Token Ratio** (lexicon) – it is the total number of unique words (types) divided by the total number of words (tokens) of a text. This metric can, then, indicate how rich the vocabulary of a text is. The closer the ratio is to 1, the more varied the vocabulary is and, consequently, more complex a text is likely to be.

   A study by Finatto (2011) has shown that scientific texts (including Biology and Pediatrics articles) tend to have a low TTR because scientists tend to repeat words consistently. Conversely, fake news articles, tend to have a richer vocabulary (FINATTO; SILVA; ESTEVES, 2021).

b) **Words Before Main Verb** (syntax) – it is the number of words that appear before the main verb of a sentence. The higher the score, the higher working memory load. Interpretation of this metric would require a careful and critical analysis by a linguist in view of a number of factors. First of all, in languages that use the canonical order SVO (Subject-Verb-Object) in their sentences (i.e., English and Portuguese), the subject can naturally precede the verb. A high score on this metric could potentially indicate complexity, but could also indicate the opposite, that is, a text producer's efforts to make their texts easy to understand, as a reader's ability to identify the subject of a sentence is very important for the effective understanding of what they read (LIBERATO; FULGÊNCIO, 2010). Second, in Brazilian Portuguese, subjects can be omitted: the suffixes added to the verbs could do the job of a subject. In English, on the other hand, the use of a subject is not optional. As a consequence, this could, in theory, mean that Portuguese tends to be naturally less complex than in English, in terms of the number of words before the main verb functioning as a subject. However, a metric alone cannot indicate complexity: it must be analyzed together with the other metrics and a human perspective. Lastly, Brazilian Portuguese is characterized, among other things, by the use of prepositions to establish connections and relations between the words. A study by

Villavicencio *et al* (2006) has shown that the most common word in Brazilian Portuguese is *de* (as in *coluna de mercúrio*), whereas the most common word in English is *the*. This indicates that collocations and noun clauses in Portuguese are formed by more words than they are in English because they are joined by prepositions. As a consequence, this could signal more complexity in Portuguese than in English, if we are to look at the words before the main verb. All of these factors deserve careful consideration if one is to deem a sentence – or perhaps a language – more complex than another.

c) **Flesch Index** (readability) – it is a metric that looks at the number of sentences of a given text, the number of words per sentence and the number of syllables of each word. All these elements can indicate how complex a text can be for a reader. Its scores can be interpreted on a scale that ranges from 0 to 100: from 0 to 25 – extremely difficult (targeted to academics); 26 – 50 – difficult (for High School or College students); 51 – 75 – easy (for Middle School students) and from 76 – 100 – very easy (for Elementary School students).

A study developed by Pasqualini (2012) used some of the metrics above – including the FI – to analyze the complexity of the English translation of short stories originally written in Brazilian Portuguese. The author also analyzed how complex the short stories originally written in English were when translated into Brazilian Portuguese. Results of her study showed that the stories translated into Portuguese were far more complex than the stories translated into English. Mention must be made that although this metric is widely used for readability purposes, because of its generic nature, it sparks a number of controversies.

Our second step in this study is to investigate which language features could make a text potentially more or less rhetorically and syntactically elaborate. A detailed and careful analysis of these features could potentially signal more or less probability for text complexity.

Results and Discussion

NILC-Metrix and Coh-Metrix have provided us with useful insights for the identification of sentence patterns of Portuguese and English. Each of the following sub-sections will present and discuss the comparative data on both corpora in view of each of the metrics that have been used.

Type-Token Ratio (TTR)

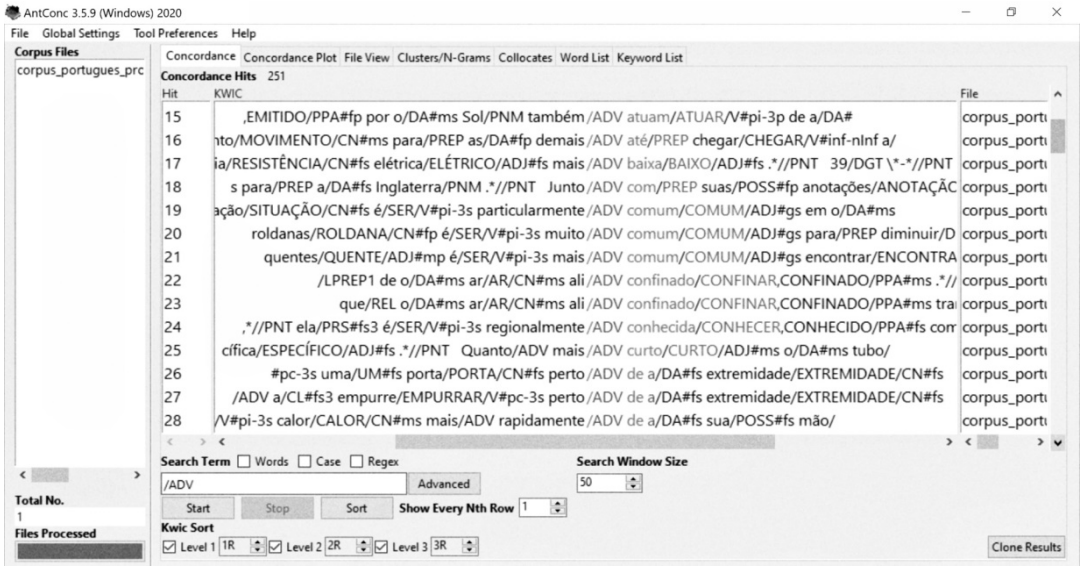
The Type-Token Ratio metric measures how rich the vocabulary presented in a text is: the more different words that are used, the more complex the text is likely to be. The TTR of the English corpus is 0.598, whereas the TTR of the Portuguese corpus is 0.677. According to the metric, the English version would be less diverse, in terms of vocabulary, than the Portuguese original. Consequently, the English version would be potentially easier.

A variety of possible features of Portuguese and English texts could have rendered the results of

the TTR. However, we have decided to look at adverbs. According to a number of studies (BECHARA, 2006; CARTER; MCCARTHY, 2006; 2010, BIBER *et al*; CECHIN; CONTINI; FINATTO, 2004; MORAES, 2015), adverbs have a reputation for adding great complexity to both Portuguese and English texts. On that line, the tagged versions of both corpora have been processed on AntConc to study the adverbs.

Figure 1 below shows the Portuguese corpus in its tagged version (using the tagline for adverbs) on AntConc.

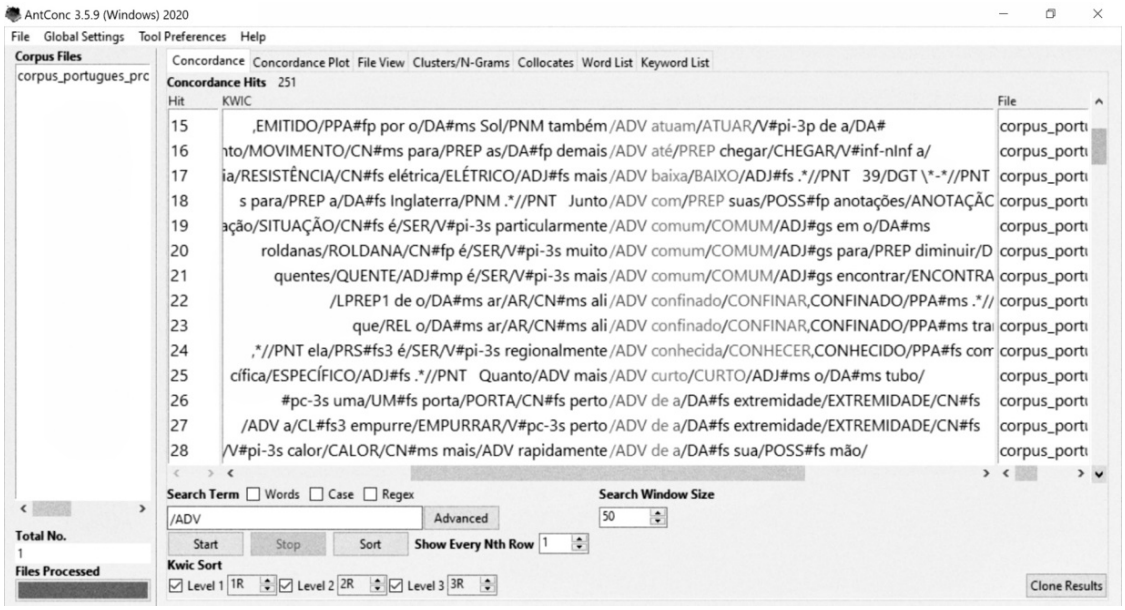
Figure 1 – Portuguese corpus on AntConc



Source: AntConc

On the other hand, Figure 2 below shows the English corpus in its tagged version (using the tagline for adverbs) on AntConc.

Figure 2 – English corpus on AntConc



Source: AntConc

AntConc delivered 251 entries for adverbs in the Portuguese corpus. On the other hand, the tool shows more adverbs in the English corpus when compared to the Portuguese corpus: 258. Some of the adverbs in the Portuguese corpus are presented in in Chart 3 below, along with a sample sentence:

Chart 3 – Portuguese Adverbs and Sample Sentences

Portuguese Adverb	Sample Sentences
Não	*Sua relativa pouca idade é a principal explicação para que a Planície Costeira **não** abrigue espécies endêmicas (exclusivas) / É mais difícil do que amarrar seus sapatos, **não**?*
Moderadamente	*Para movimentar o espelho superior do periscópio, gire **moderadamente** os cilindros laterais*
Quanto	*Observe o valor numérico da coluna de mercúrio. Quanto maior for esse valor, maior será a pressão atmosférica e vice-versa.*

Source: Authors' Own (2021)

On the other hand, Chart 4 below, presents some of the adverbs in the English corpus, along with sample sentences.

Chart 4 – English Adverbs and Sample Sentences

English Adverb	Sample Sentences
Ever	*Have you **ever** been blown away by the wind?*
More	*Liquid crystal thermometers are widely used to measure the temperature of patients because they are **more** accurate than mercury thermometers*
Officially	*During the ceremony, the winner would be **officially** announced and crowned with an olive wreath, which symbolized the utmost triumph.*

Source: Authors' Own (2021)

In the examples above, we can see the adverb *Não*, which has a moderate incidence in the Portuguese corpus. *Não* is used for a multitude of purposes. For instance, to give a possible reason why the Planície Costeira is not home for endemic species. *Não* is also used as a question tag, as a form of assertion and request for confirmation. See: *É mais difícil do que amarrar seus sapatos, não?* In this excerpt, the museum wants to know from the visitor how difficult it is to tie their shoes when using the experiment. The corpus also shows adverbs ending in *–mente* at a moderate frequency. *Moderadamente,* for example, occurs twice. In both occurrences, it is preceded by the verb *Gire*, as in an instruction to spin an object. Lastly, another relevant adverb in the corpus is *Quanto*. It occurs a number of times, mostly in a canonical structure *Quanto mais ..., maior*.

As for the English adverbs, *ever* is one whose frequency is high. MCT-PUCRS uses a pedagogical discourse to address its audiences. There are many occurrences of *Have you ever* throughout the corpus. *More* is another adverb that has a high incidence, many of which in comparative form. Lastly, adverbs ending in *–ly*, such as *officially*, occur at a moderate frequency. The number of adverbs ending in *–ly* when compared to the number of adverbs ending in *-mente* in the Portuguese corpus is marginally smaller (34 to 39).

As the data show, even though the number of adverbs on the Portuguese corpus was marginally lower than on the English corpus, the TTR of the Portuguese corpus was higher, thus indicating more elaboration and, consequently, complexity. Although many of the adverbs would appear more than once in both corpora, at least in this particular scenario, this part of speech apparently did not have a major impact on the TTR on either corpus.

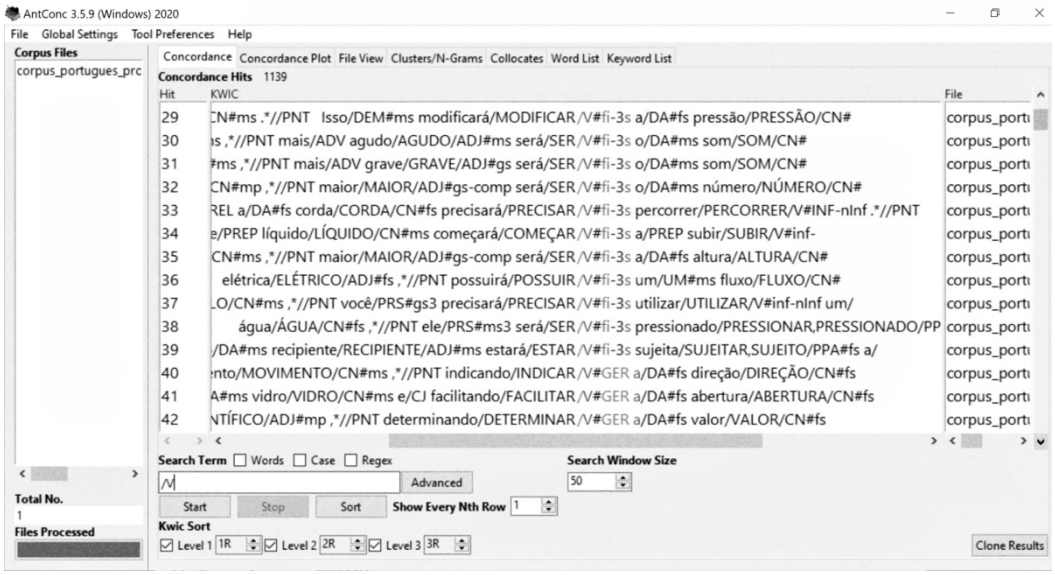
A closer examination into both corpora tried to establish the target of syntactic modification of the adverbs ended in *-ly* in English and in *-mente* in Portuguese. Adverbs of mode are known to modify verbs more frequently (BIBER *et al* 2010). The corpus showed that in both languages these adverbs largely modified verbs: in Portuguese, out of the 39, 25 presented this syntactic behavior. In English, out of the 34 adverbs that ended in *-ly*, 20 modified verbs. Therefore, there was not a deviation from the expected behavior of adverbs ending in *–ly* / - *mente* to point to adverbs as the responsible factors for the differences in the metrics. This could potentially indicate that adverbs can have a minor impact on the richness of vocabulary in Portuguese when compared to English.

Words Before Main Verb (WBMV)

The WBMV metric, as it suggests, measures the number of words that appear before the main verb of a clause. Consequently, the higher the score, the more difficult the text will be to process. The WBMV of the English corpus is 1.937, whereas the WBMV of the Portuguese corpus is 2.176. According to the metric, the English version would have fewer words, on average, than the Portuguese original. Consequently, the English version would, in theory, require less cognitive effort.

Figure 3 below shows the Portuguese corpus in its tagged version (using the tagline for verbs) on AntConc

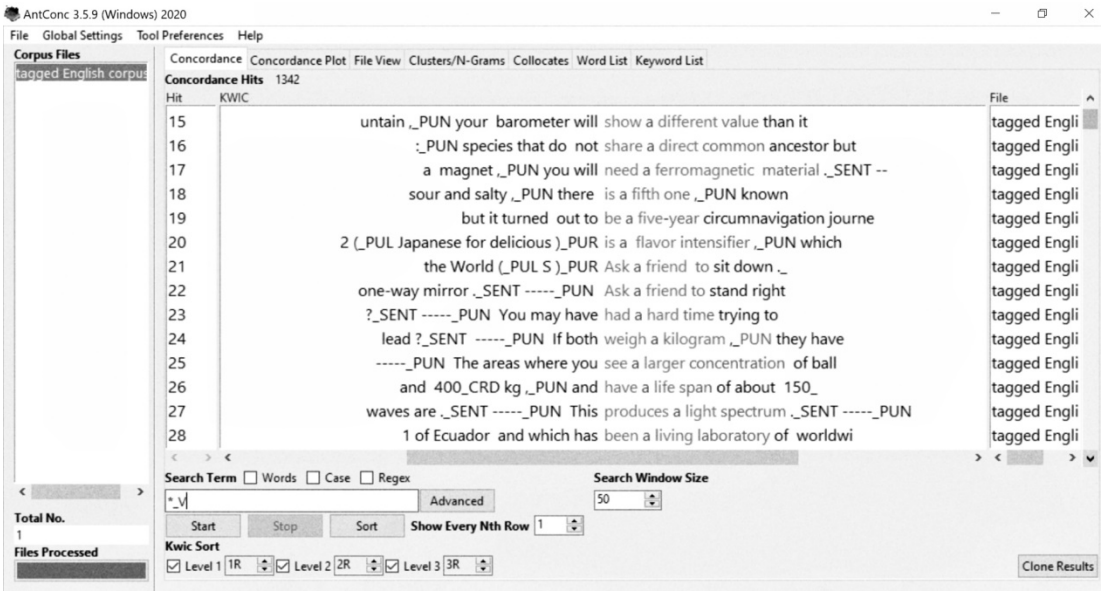
Figure 3 – Portuguese corpus on AntConc



Source: AntConc

On the other hand, Figure 4 below shows the English corpus in its tagged version (using the tagline for verbs) on AntConc.

Figure 4 – English corpus on AntConc



Source: AntConc

The tool delivered 1,139 entries for verbs in the Portuguese corpus. On the other hand, as for the English corpus, AntConc delivered 1,342 entries for verbs. When compared to its Portuguese counterpart, there are more verbs in English than in Portuguese. Some of these verbs in the Portuguese corpus are presented in the Chart 5 below, along with a sample sentence:

Chart 5 – Portuguese Verbs and Sample Sentences

Portuguese Verb	Sample Sentences
Possuir	*Mesmo um material com uma alta resistência elétrica **possuirá** um fluxo de corrente elétrica (...)*
Estabelecer	*Mesmo sem ter ultrapassado a União Soviética na Corrida Espacial, o Projeto Mercury **estabeleceu** as condições para que, alguns anos depois, o homem chegasse à Lua.*
Posicionar	***Posicione** os cilindros na extremidade da rampa e levante-a devagar*

Source: Authors' Own (2021)

On the other hand, Chart 6 below, presents some of the adverbs in the English corpus, along with sample sentences.

Chart 6 – English Verbs and Sample Sentences

English Verb	Sample Sentences
Produce	*This produces enough electrical energy to make the clock work*
Separate	*This type of mirror separates two rooms*
Pay	*Pay attention to the different sounds that are produced*

Source: Authors' Own (2021)

In the examples above, we can see some of the main verbs being preceded by many words. *Possuir*, for instance, is preceded by a phrase consisting of a sequence of 8 words. On the other hand, *estabelecer* is preceded by a 13-word phrase. Lastly, *posicionar*, which has a high incidence in the Portuguese corpus, is used mostly to give instructions. As MCT-PUCRS's directions and instructions to visitors tend to be straight-to-the-point. There are no words preceding this verb in that instruction.

As for the English verbs, *produce* occurs a number of times throughout the corpus. A structure that is somehow representative of its use is one in which it is preceded by a demonstrative pronoun taking the place of a noun phrase that was used in the previous sentence, e.g. *This*. Some other main verbs are preceded by longer chunks of language, such as *separate*. In the example above, it is preceded by a 4-word construction. Lastly, *pay*, which has a high incidence on the English corpus, is used mostly to give instructions. Just like in the Portuguese corpus, directions and instructions to visitors tend to be straight-to-the-point. There are no words preceding this verb in that instruction.

In view of the above, we can see that the number of words before the main verb is higher in the Portuguese corpus than it is in the English corpus, albeit there are more verbs in English than in Portuguese. Although some of these verbs are preceded by a variety of structures: a pronoun (be them demonstrative or of any sort), phrases (be them noun, adjectival, adverbial or of any sort), or even nothing ()), we can see that the phrases in Portuguese corpus are longer than those in the English corpus because of the number of words in them. Because they are longer, they would render Portuguese to be more rhetorically and syntactically elaborate than English, and consequently, tend to be more difficult to process.

Flesch Index (FI)

As mentioned earlier, the Flesch Index is a metric that could potentially indicate how difficult a text would be for a given reader. The FI of the English corpus is 71.350, whereas the FI of the Portuguese corpus is 52.582. According to the description of the metric, both texts would be easy for Middle School students, the English version being potentially easier.

We know that the FI does not "go alone", as a number of other factors can influence how high the score for that metric is. These factors include, among others, the number of paragraphs in a text, the number of sentences in a paragraph and the number of syllables of the words in a paragraph. Chart 7 below shows a comparison of two sample texts on our corpora to illustrate the FI metrics.

If we look at a sample below, we can see that some of the sentences of the original Portuguese were broken into two or more in the English version. See: Portuguese – *A peça pintada de vermelho é um imã permanente enquanto que a pequena lâmina verde suspensa é feita de ferro.* English – *Right behind it there is a permanent magnet in red. But the small green bar is made of iron;* This could indicate an effort of the museum to make the text more easily accessible to the reader as it is known that

longer sentences take higher cognitive load to process. Similarly, the use of *right behind it* in the English version, could be seen as a strategy to make the text easier to understand, as the museum is making use of demonstratives and adverbs to indicate the position of items in the experiment.

In what follows, the final remarks are presented.

Chart 7 – Comparison of Portuguese and English Corpora

Sample Text in Portuguese (IF: 56.495)	**Sample text in English (IF: 82.137)**
Cortando as Linhas de Indução	***Cutting induction lines***
Observe o disco do experimento. Uma de suas metades é constituída de alumínio e a outra, de ferro.	*In this experiment, you will see a yellow / blue disk. One of its halves is made of aluminum, and the other of iron.*
A peça pintada de vermelho é um imã permanente enquanto que a pequena lâmina verde suspensa é feita de ferro.	*Right behind it there is a permanent magnet in red. But the small green bar is made of iron.*
Posicione uma das metades do disco próxima à lâmina. Depois posicione a outra metade.	*Bring one of the halves of the disk next to the small green bar. After that, bring the other half closer to it.*
Por que lâmina verde não é atraída pelo imã quando está próxima ao ferro?	*Why does the magnet not attract the green bar when it is close to iron?*
Porque, diferentemente do alumínio, o ferro é um material capaz de cortar as linhas de indução do campo magnético em que está inserido.	*Because iron can cut the induction lines of its magnetic field. Aluminum cannot.*
E o que isso tem a ver com a sua vida?	***And what does this have to do with your life?***
Esse conhecimento é indispensável para todas as atividades que envolvem a utilização de imãs. Se você precisar de algo que interaja com o imã, por exemplo, você precisará utilizar um material ferromagnético.	*You should know that for anything you use magnets. If you need something to interact with a magnet, you will need a ferromagnetic material.*

Source: Authors' own (2021)

Final Remarks

The aim of this article was to study some language features that could potentially render a language as more rhetorically and syntactically elaborate than another – in the case Portuguese and English. The idea is that the more rhetorically and syntactically elaborate a language is, more complex it will be. Hence, a corpus of 100 texts (50 of them in Portuguese and 50 in English) from PUCRS' Science and Technology Museum has been lifted. Two NLP tools have been employed: NILC-Metrix and Coh-Metrix.

Results have shown that some parts of speech, such as adverbs, do not have a major impact on the richness of vocabulary, despite having reputation for adding complexity to texts, as a whole. Other language features, such as words before main verb (including anaphoric references, phrases of any sort, clauses of any sort, etc.), apparently can potentially render a language more complex than another, even when the use of words before main verbs would be necessary (for instance, when they function as subjects). Future studies, however, could be undertaken to study complexity in more depth, as this is a term whose concept is, as its name suggests, complex.

The tools that were used in our study also show their usefulness to reveal information about language complexity. In future studies, other NILC-Metrix metrics could be used, such as the Dale-Chall readability formula, which scores measures of texts in view of the words that may be considered difficult to readers. Also, a variety of different genres could be used to explore other parts of speech and possible reasons that could render one language – or genre – more or less syntactically and rhetorically elaborate than another.

References

ANTHONY, Lawrence. AntConc: A Learner and Classroom Friendly, Multi-Platform Corpus Analysis Toolkit. **Proceedings of IWLeL 2004: An Interactive Workshop on Language e-Learning**, p. 7-13, 2005.

ANTHONY, Lawrence. A critical look at software tools in Corpus Linguistics. **Linguistic Research** n. 30, v.2, p.141-161, 2013.

BECHARA, Evanildo. **Gramática escolar da língua portuguesa**. 1. ed. Rio de Janeiro: Lucerna, 2006.

BERBER-SARDINHA, Tony. **Linguística de Corpus.** Barueri, SP: Editora Manole, 2004.

BIBER, Douglas. Representativeness in Corpus Design. **Literary and Linguistic Computing**, n. 4, v. 8, p. 243–257, 1993.

BIBER *et al.* **Longman Grammar of Spoken and Written English**. Essex: Longman, 2010.

CARTER, Ronald; MCCARTHY, Michael. **Cambridge Grammar of English**: a comprehensive guide – spoken and written English grammar usage. Cambridge: Cambridge University

CECHIN, Salete Moncay; CONTINI, Daviane Zottis; FINATTO, Maria José Bocorny. Advérbios terminados em -mente (L2) e em – ly (L1): um estudo sobre condições de tradução de manuais de química. In: **Fórum Internacional de Ensino de Línguas Estrangeiras** – File III, 2004, Plotas. Anais ... Pelotas: UCPEL/UFPel, 2004.

CORTINA SILVA, Asafe David; DELGADO, Heloísa Orsi Koch; FINATTO, Maria José Bocorny. Acessibilidade Textual e Terminológica para o português brasileiro: pesquisa, estratégias e orientações de [re]escrita simplificada. **Moara**, v. 58, p. 322-343, 2021. DOI: http://dx.doi.org/10.18542/moara.v0i58.10903

FINATTO, Maria José Bocorny. **Trabalho com pequenas e grandes amostras textuais**: levantamento de terminologias na área de pneumopatias ocupacionais. In: Aparecida Negri Isquerdo; Giselle Olívia Mantovanni Dal Corno. (Org.). As Ciências do Léxico: Lexicologia, Lexicografia, Terminologia. Volume VIII. 1 ed. Campo Grande, MS: Editora da Universidade Federal do Mato Grosso do Sul Ed. UFMS, 2018, v. 8, p. 347-372.

FINATTO, Maria José Bocorny; MOTTA, Ester. Terminologia e Acessibilidade: novas demandas e frentes de pesquisa. **Revista GTLex**, v. 2, n. 2, pp. 316-356, jan. 2019. DOI: https://doi.org/10.14393/lex

FINATTO, Maria José Bocorny; TCACENCO, Lucas Meireles. Tradução intralinguística, estratégias de equivalência e acessibilidade textual e terminológica. **Tradterm**, 37 (1), p. 30-63, 2021; Disponível em: https://www.revistas.usp.br/tradterm/article/view/168327; Data de Acesso: 02 jun 2021. DOI: https://doi.org/10.11606/issn.2317-9511.v37p30-63.

FINATTO, Maria José Bocorny. Complexidade textual em artigos científicos: contribuições para o estudo do texto científico em português. **Organon** (UFRGS), v. 50, p. 30-45, 2011.

FINATTO, Maria José Bocorny; SILVA, Adriana da; Esteves, Francine Facchin. Fake news e desinformação sobre vacinas: contribuições dos estudos da Terminologia, do Texto e do Discurso. **Revista GTLex**, v. 6, p. 445-494, 2021. DOI: 10.14393/lex

GRAESSER, Arthur. *et al*. Coh-Metrix: Analysis of text on cohesion and language. **Behavioral Research Methods**, p. 193-202, 2004.

GRIES, Stefan. Corpus linguistics and theoretical linguistics. A love–hate relationship? Not necessarily... **International Journal of Corpus Linguistics**, n. 15, v.3, p. 327-343, 2010. DOI: 10.1075/ijcl.15.3.02gri

JAKOBSON, Roman. On linguistic aspects of translation. In: Venuti, L. **The Translation Studies Reader**. London: Routledge, 113-118, 2000 [1959].

LEAL, S. M.; DURAN, M. S.; SCARTON, C. E.; HARTMANN, N. S.; ALUÍSIO, S. M. 2022. NILC-**Metrix**: assessing the complexity of written and spoken language in Brazilian Portuguese. CoRR abs/2201.03445 (2022) https://arxiv.org/abs/2201.03445. DOI: https://doi.org/10.48550/arXiv.2201.03445

LIBERATO, Yara. (Org.); FULGÊNCIO, Lúcia. (Org.). **É possível facilitar a leitura - um guia para escrever claro**. São Paulo: Contexto, 2007. 175p.

MCENERY, Tony; HARDIE, Andrew. **Corpus linguistics**. Cambridge: Cambridge University Press, 2012.

MORAES, Helmara Febeliana Real de. A Questão da Equivalência entre os Advérbios em -LY e -Mente no Inglês e no Português – como funciona em linguagens especializadas? In: TAGNIN, Stella E.O.; VIANA, Vander. (Org.). **Corpora na Tradução**. 1 ed. São Paulo: Hub Editorial, 2015, v., p. 105-130, 2015.

NÚCLEO INTERINSTITUCIONAL DE LINGUÍSTICA COMPUTACIONAL. **Coh- Metrix-Port**. Versão 3.0. São Paulo: Universidade de São Paulo, NILC, 2020.

PASQUALINI, Bianca Franco. **Leitura, Tradução e Medidas de Complexidade Textual em Contos da Literatura para Leitores com Letramento Básico**. 155 p. Dissertação (Mestrado em Letras). Instituto de Letras, Universidade Federal do Rio Grande do Sul. Porto Alegre – RS, 2012.

PUCRS. Como nascem as exposições do museu? **Revista da PUCRS**, Porto Alegre, edição 190, p. 40-42, jul-set. 2019. Disponível em: https://www.pucrs.br/revista/wp-content/uploads/sites/136/2019/06/revista_pucrs-0190.pdf.

RODRIGUES, Erica dos Santos; FREITAS, Cláudia; QUENTAL, Violeta. Análise de inteligibilidade textual por meio de ferramentas de processamento automático do português: avaliação da Coleção Literatura para Todos. **Letras de Hoje** (Impresso), v. 48, p. 91-99, 2013.

SCARTON, Carolina Evaristo; ALUÍSIO, Sandra Maria. Análise da Inteligibilidade de textos via ferramentas de Processamento de Língua Natural: adaptando as métricas do Coh-Metrix para o Português. **Linguamática** (Revista para o Processamento Automático das Línguas Ibéricas), v.2, n.1, p.45-61, 2010. Disponível em: http://linguamatica.com/index.php/linguamatica/article/viewfile/44/59. Acesso em: 20 set. 2021.

SINCLAIR, John. **Corpus, Concordance, Collocation: Describing English Language**. Oxford: Oxford University Press, 1991.

TCACENCO, Lucas Meireles. RODRIGUES DA SILVA, Bruna. FINATTO, Maria José Bocorny. Acessibilidade textual e terminológica: conquistas recentes e novas frentes de pesquisa. **Revista GTLex**, n.*3, v.*2, p. 197-224, 2019. DOI: 10.14393/Lex6-v3n2a2018-1.

VIANA, Vander; TAGNIN, Stella Esther Ortweiler. (org.). **Corpora na Tradução**. São Paulo: Hub Editorial, 2015.

VILLAVICENCIO, Aline; FINATTO, Maria José Bocorny; POSSAMAI, Viviane. Padrões da Preposição DE entre sintagmas nominais em linguagem cotidiana e linguagens técnico-científicas. **V Encontro de Corpora**. São Carlos – SP, v. 01, p.01, 2006.

WEIGAND, Edda. Possibilities and Limitations of Corpus Linguistics. In: **Dialogue Analysis VIII**: Understanding and Misunderstanding in Dialogue: Selected Papers from the 8th IADA Conference, Goteborg 2001, edited by Karin Aijmer, Berlim, New York: Max Niemer Verlag, p. 301-318, 2011.

ZETHSEN, Karen Korning. Intralingual translation: an attempt at description. **Meta**, Montreal – Canadá, 54 (4), pp. 795-812, 2009; DOI: https://doi.org/10.7202/038904ar; Data de Acesso: 14 mai 2021.