DOI: 10.14393/DLv17a2023-53
e1753

Article

A natural language processing approach to complexity assessment of 18th-century health literature

Uma abordagem de processamento de linguagem natural para avaliação de complexidade em literatura médica do século XVIII

Leonardo ZILIO[*]
Maria José Bocorny FINATTO[**]
Renata VIEIRA[***]
Paulo QUARESMA[****]

ABSTRACT: In this paper, we present an experiment for complexity-level analysis of Portuguese texts from the 18th century using NLP tools. The 18th century was the time for the realization of a new world that had been built since the Renaissance, it was the period of consolidation of many of the current sciences. One of its characteristics is the presentation of scientific written records in national languages, rather than Latin, and the expressed wishes that the specialized texts could be more understandable to people of lesser erudition. As such, we intend to collaborate to identify if and how these wishes were fulfilled. To achieve this goal, we resort to an NLP supporting methodology to detect degrees of complexity of a medical work of this time period and compare it with two other works that have hypothesized lesser and greater complexities. By using NILC-Metrix, we intend to identify features of a continuum of complexity in this kind of document.
KEYWORDS: Textual complexity. 18th-century Portuguese. Historical Linguistics. Historical Terminology. Digital Humanities.

RESUMO: Neste artigo, apresentamos um experimento que usa ferramentas de PLN para analisar o nível de complexidade de textos em português do século XVIII. Trata-se de um período de concretização de um novo mundo, que se iniciou com a Renascença, e de consolidação de muitas das ciências modernas. Nessa época, também começaram a surgir publicações científicas em línguas vernaculares, e não em latim, uma decorrência da vontade de tornar textos especializados mais acessíveis a pessoas de menor erudição. Nesse contexto, nosso objetivo é tentar identificar se, e como, esses ideais de acessibilidade foram atingidos. Para tal, nos apoiamos numa metodologia de PLN para detectar níveis de complexidade de uma obra médica desse período, comparando-a com outras duas obras de complexidade

[*] Post-doctoral Researcher. Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Germany. leonardo.zilio@fau.de
[**] Full Professor of Linguistics. Universidade Federal do Rio Grande do Sul (UFRGS), Brazil. mariafinatto@gmail.com
[***] Principal Investigator. CIDEHUS, Universidade de Évora, Portugal. renatav@uevora.pt
[****] Full Professor of Informatics. Universidade de Évora, Portugal. pq@uevora.pt

hipotética maior e menor. Usando a ferramenta NILC-Metrix, nosso objetivo é identificar um contínuo de complexidade nesses documentos.
PALAVRAS-CHAVE: Complexidade Textual. Português do século XVIII. Linguística Histórica. Terminologia Histórica. Humanidades Digitais.

Article received: 06.26.2023
Article approved: 10.27.2023

1 Introduction

Information extraction from corpora is an increasingly relevant Natural Language Processing (NLP) task (Piotrowski, 2012). Its main objective is to automatically structure and represent knowledge from dispersed documents. Document collections composed of historical texts, printed or handwritten, have an enormous information potential, but they are usually not easily accessible to researchers or citizens in general. And, even when available in digitized format, reading comprehension of these old materials, for different reasons, is not a trivial task, as has been shown in Finatto (2018), Finatto (2020), and Quaresma and Finatto (2020), where NLP tools were applied to automatically analyze documents from the 18th century and to extract and represent the associated information.

In this context, Finatto, Quaresma and Gonçalves (2018) developed a research project with the purpose of describing and systematizing the content and the lexicon-grammatical features of a series of printed works on Medicine, published throughout the 18th century. The 18th century was the time for the realization of a new world that had been built since the Renaissance. It was a period of consolidation for many of the current sciences (Barbosa, 2020). One of its characteristics is the presentation of scientific written records in the national languages, rather than Latin, and the expressed wishes that specialized texts could be more understandable to people of lesser erudition (Finatto, 2018). These ideals of making texts more accessible are in line with a Lutheran heritage of access to information according to people's needs (Lobenstein-Reichmann, 2022). As such, we intend to identify if and how these wishes

were fulfilled. To achieve this goal, we use an NLP methodology to detect the probable degrees of complexity of a medical text from this time period and compare it with works of hypothesized lesser and greater complexities. The medical text that we used was extracted from the work *Observaçoens medicas doutrinaes de cem casos gravissimos* [free translation: *Medical and doctrinal observations of a hundred severe cases*] by João Curvo Semedo (1635-1710). This text was compared with an excerpt from the *Gazetas Manuscritas* (Vol. I, 1729-1731) [free translation: *Handwritten News*] and a sermon by Fr. Antônio Vieira (1608-1697). These three works are presented in detail in Section 3. The application of this methodology is a novel approach for studying the complexity of historical Portuguese texts.

Our initial hypothesis is that Semedo's medical writing would be, on average, more complex than the less formal writing style of the *Gazetas Manuscritas*, which contain texts from the 18th century that were written by lay people in order to present everyday news and facts to other lay people. However, it would be less complex than the erudite *Sermons* by Fr. Antônio Vieira, which are considered representative of the Baroque and refined language. To verify this assumption, we worked with an orthographically modernized version of texts extracted from these three works. We opted to use orthographically modernized versions because of the higher precision of NLP tools using modern-day texts. We used one text from each of the works. While this number is restrictive, the task of modernizing the transcriptions' orthography is not trivial and has to be done manually, which makes it hard to carry out on a large scale. As such, we deemed that the selected texts are enough for an initial, exploratory research into the complexity of historical texts.

We used the NILC-Metrix system (Gazzola; Leal; Aluisio, 2019) to help us identify features of a continuum of complexity in the documents. This system is, as far as we know, the most complete tool for Portuguese complexity analysis. It brings together metrics that were developed for over more than a decade in the Interinstitutional Center for Computational Linguistics (NILC), starting with Coh-

Metrix-Port (Caseli *et al.*, 2009; Aluísio; Gasperin, 2010; Cunha, 2015), an adaptation of the original English Coh-Metrix (Graesser *et al.*, 2004) to Brazilian Portuguese. The main purpose of these metrics is to assess text complexity level. Today NILC-Metrix comprises 200 metrics. The system was designed for general-purpose, non-scientific texts and, although being from different domains, our selected texts are also non-scientific, as they were meant to be understandable by lay people. NILC-Metrix was developed for modern texts and, as far as we know, our experiments are the first ones to apply it to texts from the 18th century. As such, a deeper linguistic analysis of the application of these measures to texts from the 18th century would be required for achieving more thorough conclusions. However, as we are not comparing text complexities between different time periods, a comparative evaluation using NILC-Metrix is a sound way of gathering information about relative text complexity.

In the next sections, we present an overview of the metrics in NILC-Metrix. Then the corpus is presented in detail, with brief details about the authors and text types that are under investigation, according to their time period, and we also provide text samples that illustrate the different writing styles. We then move on to discuss the methodology for a comparative analysis of text complexity. Finally, we present the observed results, their analysis and limitations, leading to some final remarks regarding perspectives for the expansion of this type of study.

2 Complexity metrics

The NILC-Metrix online system (Leal *et al.*, 2021) is an implementation of a set of 200 textual complexity metrics developed for over more than a decade in the Interinstitutional Center for Computational Linguistics (NILC). These 200 metrics are distributed along 14 groups, which we characterize in more detail in the following paragraphs:

1. **Descriptive measures** (10 metrics): this category covers basic counts like absolute and average number of words per sentence, number of syllables per word, and also number of sentences, paragraphs and subheadings. This first group of measures was not used in the complexity analysis because the texts used in this study were subject to specific constraints in terms of size, such as the limit of 2,000 words in each text that was submitted to the online NILC-Metrix system.
2. **Textual simplicity** (8 metrics): this group shows counts for elements such as personal pronouns, ratio of easy conjunctions to number of words in the text, ratio of difficult conjunctions to number of words in the text, and number of long sentences.
3. **Referential cohesion** (9 metrics): this category of measures covers the network of references between words, considering pronouns and expressions throughout the sentences. The referents are nouns and pronouns, which are repeated in adjacent sentences. This group also covers word repetition in adjacent sentences.
4. **Semantic cohesion** (11 metrics): this set of measures has to do with similarity between pairs of adjacent sentences throughout the text. Different metrics, dealing with Latent Semantic Analysis (LSA), compute similarity between text excerpts (words, sentences) considering implicit knowledge in addition to similar words.
5. **Psycholinguistic measures** (24 metrics): this set of measures focuses on psycholinguistic properties of words gathered in the lexical database PortLEX[1] (Santos *et al.*, 2017). As a psycholinguistic repository, it is a lexical resource with values for four psycholinguistic characteristics of words: concreteness, familiarity, age of acquisition and imageability.
6. **Lexical diversity** (15 metrics): in this group, the different proportions of quantities and varieties of words throughout the sentences and the text are observed. This group measures, for example, the quantity and variety of grammatical words in relation to the quantity of content words, such as adjectives and adverbs. These measurements quantify the total number of words and the number of words that are repeated in the text.
7. **Connectives** (12 metrics): this group of measures is based on a list of Brazilian Portuguese connectives/operators and classifies them as additive, causal, logical, conditional, *etc.*, assigning them positive or negative semantic values. The logical operators category covers words that establish logical relationships in the text, such as: ou [or], e [and], se [if], não [no].

[1] For more information about PortLEX: http://143.107.183.175:21380/portlex/index.php/en/?option=com_content&view=article&layout=edit&id=23 (accessed on: 21 Jun. 2023).

8. **Temporal lexicon** (12 metrics): this group is dedicated to quantifying the proportion of positive and negative temporal connectives – established in a list of words – as well as to considering past tense verbs throughout the text.
9. **Syntactic complexity** (27 metrics): this group of measures assesses the complexity of sentence structures. Elements such as the ratio of clauses per sentence, the average number of words before the main verbs of the main clauses, as well as the average number of adverbial adjuncts per clause or the number of coordinate conjunctions per sentence in the text are considered.
10. **Density of syntactic patterns** (4 metrics): this measurement group contains elements such as the average length of noun phrases in sentences, the ratio of gerund verb forms to all verbs in the text, and the different lengths of noun phrases throughout the text.
11. **Word-level morphosyntactic information** (42 metrics): this measure includes various elements such as the proportion of adjectives, nouns, adverbs and pronouns in relation to the number of words in the text. The number of content words and functional and/or grammatical words in relation to the total number of words in the text is also evaluated.
12. **Word-level semantic information** (11 metrics): in this group, the proportion of polysemy for different word classes is evaluated. For example, one metric checks the ratio between the number of meanings of verbs and the number of different verbs used in the text. Also in this group is the proportion of abstract nouns in relation to the number of words in the text.
13. **Word frequency** (10 metrics): in this group, the averages of word frequencies are observed. The corpus *Banco de Português* (BP) (Berber Sardinha; Barbara, 2005) in its 2010 version is used as reference for some of the metrics. The BP at the time was the largest and most balanced corpus for Brazilian Portuguese, so these metrics were kept for historical reasons. However, the new *Brazilian Corpus*[2] and *brWaC* (Wagner Filho *et al.*, 2018) are also used for frequency counts using a Zipf scale.
14. Readability indexes (5 metrics): this category includes five different indexes (Santos *et al.*, 2020) that were adapted to Brazilian Portuguese. These indexes evaluate the level of understanding and readability of the text:
    I. The *Brunet Index* is a form of type/token ratio that is less sensitive to text size. It first raises the number of types to a constant (-0.165) and then uses the result as the power to which the number of tokens is raised.
    II. The adapted *Dale Chall* readability formula combines the number of unfamiliar words with the average number of words per sentence.

[2] For more information about the Brazilian Corpus: https://www.linguateca.pt/acesso/corpus.php?corpus=CBRAS (accessed on: 21 Jun. 2023).

III. The *Flesch Index* is a readability index that seeks to find a correlation between the average sizes of words and sentences, and measures how easy a text is to read. The original equation was adapted to Brazilian Portuguese by Martins et al. (1996).
IV. The *Gunning Fog Readability Index* (also known as Gunning FoX) adds the average number of words per sentence to the percentage of difficult words in the text, and then multiplies it by 4. The result is directly linked to the 12 levels of American education. Indexes greater than 12 represent extremely complex texts.
V. The *Honoré*'s statistic is a type of type/token ratio that takes into account, in addition to the number of types and tokens, the number of *hapax legomena* (*i.e.*, words that occur only once).

The Dale Chall Formula and Gunning Fog Readability Index vocabulary references were extracted from a list of Brazilian Portuguese simple words based on a dictionary for kids (Biderman; Carvalho; Pedroso, 2004).

The NILC-Metrix system (Leal *et al.*, 2021) has an online interface[3] with a limit of 2,000 words. Therefore, we examined one text within this limit from each author in our corpus from the 18th century.

3 Corpus

This study is concerned with an extract from Semedo's medical work, the main component of our historical corpus[4]. As contrastive material, we are using Fr. Vieira's *Sermons* and a collection of handwritten newspapers called *Gazetas Manuscritas da Biblioteca Pública de Évora* [free translation: *Handwritten News of Évora's Public Library*] (see a part of this in Lisboa *et al.* [2002]). While these texts are all from a similar period, they were transcribed using different methodologies and for different ends, so, in this section, we present extracts of the texts in their original transcribed format.

[3] The online interface for NILC-Metrix is available at: http://fw.nilc.icmc.usp.br:23380/nilcmetrix-en (accessed on: 21 Jun. 2023).
[4] The corpus is partially available at https://sites.google.com/view/projeto38597?pli=1 and https://www.ufrgs.br/textecc/terminologia/page/index/ (accessed on: 27 Oct. 2023).

3.1 Curvo Semedo's work

Our starting point is the handbook *Observaçoens medicas doutrinaes de cem casos gravissimos* [free translation: *Medical and doctrinal observations of a hundred severe cases*] (Semedo, 1707). It was printed in Lisbon, in 1707, and contains 635 pages. The handbook was written by João Curvo Semedo (1635-1719), a Portuguese physician from Monforte, Alentejo, a region in Portugal. Semedo produced several medical treatises and handbooks, such as the *Polyanthea medicinal* (1697) and the *Atalaya da vida contra as emboscadas da morte* [free translation: *Observatory of life against the traps of death*] (1720). These two books, among others in Semedo's records, also have more than 600 pages. Due to his extensive bibliographic production, Semedo was one of the "most popular doctors throughout the Portuguese empire in the eighteenth century" (Furtado, 2008, p. 147). In addition to some well-known and manipulated chemical substances, some innovative treatments prescribed by Semedo, called "the Curvian secrets", were made with ingredients from Brazil, Africa and Asia. Semedo's new authorial treatments – some very bizarre by today's standards – are always highlighted in his books. They indicate that European medicine was open to using products from other regions of the world.

It is important to emphasize that Semedo's proposal intended to present these texts, vocabularies and terminologies in a way that would make them accessible to the reader, with special attention to the lower literate "young doctors" of his time, who did not know enough Latin but could read a text in Portuguese. However, in linguistic terms, his work represents the period of the "classical Portuguese" (Castro, 2006; Banza; Gonçalves, 2018) and illustrates the medical terminology of this period. Although the emergence of Portuguese language terminologies represents a true technological metamorphosis of the language (Verdelho, 1998), their historical analysis still lacks a systematic study, which is also true for medical terminology. As an illustration of Semedo's writing style, we present the following extract reporting a case of renal colic, which preserves the original spelling:

OBSERVAÇAM XLV.

De hum mercador , a quem repentinamente assaltou huma dor de colica taõ intoleravel , que estando na fé sacramental para commungar , o naõ pode fazer ; e sendo eu chamado , conheci dos grandissimos ardores ; e continuos desejos de ourinar , e vomitar , das picadas da bexiga , e do adormecimento da perna direita , que a tal dor era nephritica ; para cujo remedio appliquei hum vomitorio de tres onças de agua benedicta vigorada , e tres ajudas feitas de cozimento de rim de vacca , misturando em cada huma , huma onça de terebinthina de beta lavada em agua de malvas , atè se fazer muito branca , com huma gema de ovo crua ; metendo depois disto ao doente em hum banho de agua , q primeiro fosse cozida com hum arratel de amendoas doces bem pizadas ; e foraõ estes remedios taõ maravilhosamente succedidos , que dentro de quatro horas deitou muitas pedras redondas do tamanho de grãos de pimenta , e no mesmo dia se tiràraõ as dores , e ficou sao .

3.2 *As Gazetas Manuscritas da Biblioteca Pública de Évora*

The collection *As Gazetas Manuscritas da Biblioteca Pública de Évora* [free translation: *The Handwritten News of Évora's Public Library*] (Lisboa; Miranda; Olival, 2002) is a large corpus of journalistic texts from the 18th century. It was written in Portuguese by lay people in order to present everyday news and facts to other lay people. This collection is partially available – in a transcribed version – with the Tycho Brahe Corpus (Sousa, 2014)[5]. As an illustration of this early newspaper writing style, we present the excerpt below, which contains a modernized spelling:

Diário de 23 de agosto de 1729

Pelas cartas de Vasco Fernandes César, se soube a notícia, que aqui todos ignoravam, de que El-rei o tinhafeito Conde de Sabugosa vila junto a Viseu de que não sabemos se lhe desse senhorio.

Chegou Rodrigo César, gordo, mas não cheio, mostrou grande desinteresse; as minas que descobriu temgrande quantidade de ouro, e se achou um grão de meia arroba, porém é mau o clima, e tão dilatado o caminho, que comeu várias coisas asquerosas, ficou de posse do governo Antônio da Silva Pimentel.

[5] The Tycho Brahe Corpus is available at: http://www.tycho.iel.unicamp.br/corpus/ (accessed on: 21 Jun. 2023).

> Elegeu-se abadessa de Santa Clara Dona Maurícia, de muita capacidade tendo praticado Frei Antônio da Piedade a renúncia voluntária da sua antecessora, de que se espera o sossego daquele convento.
> Chegou a nau de Macau com João Baptista Rollano, e a vida da Índia, por onde se sabe que sem embargo da conquista de Mombaça e todas as suas grandes dependências, não faltava cuidado, porque os arábios, diziam estavam unidos, e cuidavam em restaurá-la, e os reis vizinhos tem má inteligência, fazendo pelo Norte algumas entradas; a nau que de Lisboa partiu em abril arribou à Bahia, de onde dizem irá com outra, que prepara o Conde de Sabugosa, e aqui se apressam as que vão a Mombaça.

In relation to the above-mentioned medical books, the *Gazetas Manuscritas* can be considered as an initial contrastive reference of textual and linguistic simplicity. It works as a typical record of day-to-day topics from that time, something related to the everyday common language, dealing with non-specialized subjects.

3.3 Fr. Antônio Vieira's *Sermons*

Fr. Antônio Vieira (1608-1697) was born in Portugal. His bibliographic work is known to this day for the complexity of the text and the sophisticated rhetoric. In 1614, he and his family moved to Brazil, where he studied at a Jesuit college and later joined the Society of Jesus. In 1633 he delivered his first two sermons, both political in character: the first one spoke of the Dutch invasion, and the second one attacked indigenous slavery. His political views, coupled with his defense of the New Christians, generated enmities among the settlers and members of the Church.

Fr. Antônio Vieira's Sermons are part of the Baroque language and have, among others, the following characteristics: a cult of contrast, refined language, the use of antitheses and paradoxes. In addition to sermons, Vieira also wrote letters, prophetic texts, poetry, and theater. His life and work were marked by his involvement in political issues, in defense of the Portuguese Crown and the Catholic faith (in spite of the fact that Vieira was persecuted by the Portuguese Inquisition). After spending some time in Portugal and in Italy, he died in Brazil in 1697.

A collection of his sermons is partially available – in a transcribed version – with the Tycho Brahe Corpus (Sousa, 2014). In order to illustrate his writings, we present the following extract, which contains a partially modernized spelling:

> **SERMÃO | da | Primeira Dominga do Advento**
> **Prégado na Capella Real, no Anno de 1652 | Amen dico vobis, non praeteribit | generatio haec donec omnia fiant. | Lucas, XXI | I**
>
> Muitas coisas sabemos deste grande dia, todas grandes e temerosas, e duas só ignoramos. Sabemos que antes do dia do Juízo , o sol, que soía fazer o dia, se há-de escurecer e esconder totalmente com o mais horrendo e assombroso eclipse que nunca viram os mortaes. Sabemos que a lua, não por interposição da terra, mas contra toda a ordem da natureza, se há-de mostrar entre as trevas medonhamente desfigurada, e toda coberta de sangue. Sabemos que as estrellas do firmamento, desencaixadas das orbes celestes, hão de cahir: e como no mundo inferior não têm onde caber, lá hão de estalar a pedaços, com horrível estrondo, e exhalar-se em vapores ardentes. Sabemos que o mar há-de sahir furiosamente de si, e atroar os ouvidos atônitos com pavorosos roncos, e levantando ondas immensas até às nuvens, já não há-de bater como dantes as praias; mas sorver inteiras as ilhas, e afogar os montes.

3.4 Spelling standardization and descriptive measures

The three excerpts presented in the previous section are a reflex of different spelling choices in the transcription of the original texts. Semedo's excerpt preserves the original spelling, with words such as *tiràraõ* (instead of the modern *tiraram* [removed$_{\text{3rd person plural}}$]). Vieira's sermon presents words using the modern spelling, but others still in an archaic form, such as *cahir* (instead of *cair* [to fall]). The Gazetas, however, are fully modernized, but still present a few words that were not tokenized (they are not separated by spaces), such as *tinhafeito* (instead of *tinha feito* [had made$_{\text{1st or 3rd person singular}}$]). These different spellings have an impact on the automatic processing tools that are applied, and so the results of the metrics that we use would not be comparable if we were to use these three different spellings.

To avoid this issue, and to make sure that we have comparable samples and results, we opted for modernizing all three excerpts, so that all have the same spelling as presented in the excerpt from the *Gazetas*. This modernization was done manually by one of the authors, by first converting the texts to a plain text format, and then using Trados Studio 2021, a computer-assisted translation (CAT) tool, as an interface to make the changes in spelling[6]. This resulted in a very laborious process that, unfortunately, would be hard to apply to a larger corpus without the help of several trained language specialists. Table 1 presents some descriptive information from NILC-Metrix about each of the three texts in their modernized form, after the standardization process was applied.

Table 1 – Corpus description.

	Gazetas	Semedo	Vieira
Tokens	1845	2016	1803
Sentences	48	25	41
Paragraphs	48	15	6

Source: created by the authors.

4 Methodology for a comparative analysis

The goal of this study is to assess the textual complexity in texts from three authors under the hypothesis that the ranking from the less to the more complex text would be Gazetas < Semedo < Vieira. In order to achieve this goal, we observed the results in a comparative way, that is, the texts were analyzed in relation to each other.

Since each metric has its own scale, for conducting a global analysis, we first had to identify those metrics in which higher values in the output corresponded to higher complexity, and those in which higher values meant lower complexity. However, not all metrics fall in these two categories. For instance, several measures that rely on the standard deviation of a specific metric were not relevant for us, as they

[6] The modernized versions of the texts used in this study are available as plain text files at: https://github.com/uebelsetzer/complexity_18th-century_texts_PT.

cannot be directly used to gauge whether a text is more complex than another. This individual analysis to attribute a specific category of complexity to each of the metrics resulted in 51 metrics not being used in the results that are presented in the next section[7].

After this categorization of the metrics, each text was ranked as 1st, 2nd and 3rd, where the first was the more complex, and the third was the less complex, according to how the metric should be interpreted. For a few metrics, there were ties between all the three texts or between two of them. In these cases of tie, we did not penalize the texts, and we only attributed either Rank 3 for all (all were considered less complex), or Ranks 2 and 3.

In order to combine these counts into one comparable value, for each text we multiplied the number of Rank 1 occurrences by 3 and the number of Rank 2 occurrences by 2, and then added their sum to the number of Rank 3 occurrences. So, the lower the rank, the more weight it had towards complexity. The analysis of complexity was then conducted in a global scale, using all 149 metrics (excluding the 51 that were removed in the process explained above), but also in more focused scales, as it is discussed in the next subsection. With this procedure, we were able to have not only a general comparative evaluation, but also a specific and detailed analysis of the main classes of measures (presented in Section 2).

4.1 Categories of analysis

The 149 metrics that were used for analyzing the texts do not include descriptive measures (Category 1 in Section 2), because the texts that we selected respected a few constraints, such as having around 2,000 words, and so these descriptive metrics would not be a good fit for comparing them.

[7] A full list indicating metrics that were used in this study is available as a CSV file at: https://github.com/uebelsetzer/complexity_18th-century_texts_PT.

We have also discarded from the experiment any metric that did not have a clear indication, expressed in NILC-Metrix's user guide, of how it directly contributes to evaluating text complexity. This was the case, for example, of measures that indicated maximum and minimum amount of content words in a sentence, and measures that involved standard deviation. We opted for discarding these metrics because measures that point to a maximum and a minimum could be directly affected by errors in sentence splitting. Other discarded cases were those measures that seemed redundant or not very well adapted to consider Portuguese grammatical features, such as the proportion of inflected verbs. In these examples, the greater or lesser presence of conjugated verbs by itself would not correspond to greater or lesser complexity given that the "default" in Portuguese is the use of inflected verbs compared to the occurrence of non-inflected forms of the verbs.

However, this general complexity analysis using 149 metrics merges and, therefore, also hides the different linguistic aspects of the texts, so some of the metrics are also shown separated into more fine-grained classes of cohesion complexity, syntactic complexity, readability, psycholinguistic complexity and lexical complexity. The following groups, which combine different sets of metrics, were established by comparing their nature and scope, and are further elaborated here:

- **Cohesion complexity analysis** (15 metrics): Referential cohesion and Semantic cohesion (groups 3 and 4).
- **Syntax complexity analysis** (40 metrics): Connectives, Syntactic complexity, Density of syntactic patterns (groups 7, 9, and 10).
- **Readability complexity analysis** (5 metrics): this included only the readability metrics (group 14).
- **Psycholinguistic complexity analysis** (20 metrics): this group was formed only by the psycholinguistic metrics (group 5).
- **Lexical complexity analysis** (63 metrics): Lexical diversity, Temporal lexicon, Word morphosyntactic information, Semantic word information e Frequency of words (groups 6, 8, 11, 12, and 13).

5 Results and discussion

In this section, we present the global analysis that was generated based on the 149 selected metrics from NILC-Metrix, as well as on the more fine-grained analysis divided by classes of linguistic features[8]. We also briefly discuss each of the findings in relation to our initial hypothesis.

Table 2 presents the overall rank of each text, where the global value was reached by the following equation, as explained in Section 4:

$$\textbf{Global} = (\textbf{Rank 1} \times 3) + (\textbf{Rank 2} \times 2) + (\textbf{Rank 3} \times 1)$$

In this overall ranking the text by Semedo was ranked as the most complex text of the three samples, with a global score of 313, just slightly above Vieira. The results from this general analysis go partially against our hypothesis that Semedo's text would be simpler than Vieira's, but it does confirm that the text from the Gazetas is the least complex among the three.

Table 2 – General complexity analysis (149 metrics).

	Gazetas	Semedo	Vieira
Rank 1 High	38	55	51
Rank 2 Med	37	54	59
Rank 3 Low	74	40	39
Global	**262**	**313**	**310**

Source: created by the authors.

Besides this general analysis, we also broke up the results in several categories, which correspond to different linguistic features of the texts. These more fine-grained categories present a more focused analysis of the different textual aspects that

[8] A complete table with all metrics and the full analysis conducted in this study is available at: https://github.com/uebelsetzer/complexity_18th-century_texts_PT. A CSV file contains a score for each text and each metric, showing also whether or not the metric was used in this study, as discussed in Section 4.

contribute to complexity, and we also use them to evaluate whether these fine-grained features align with our hypothesis.

Table 3 shows the cohesion complexity of the three texts, which considers similarity between sentences and complexity of references inside the texts. It encompasses a total of 13 metrics. All three texts have similar scores, and this is the only group of metrics where Vieira is considered the least complex. This potentially means that the text of Vieira has a more fluid transition between sentences and paragraphs when compared to the other two samples, while Semedo is again the most complex to read in this regard, albeit not by much. The text from the Gazetas presents very short news items one after the other, where the relation between paragraphs is normally not present (because it usually is a new news story), and this might explain why it scores higher than Vieira in this category.

Table 3 – Cohesion complexity analysis (15 metrics).

	Gazetas	Semedo	Vieira
Rank 1 High	3	7	5
Rank 2 Med	9	2	4
Rank 3 Low	3	6	6
Global	**30**	**31**	**29**

Source: created by the authors.

Table 4 presents the syntactic complexity of the texts, which considers, for instance, the use of connectors, phrase sizes and use of passive voice. Here the score of Semedo's text lies in the middle, while Vieira has a higher score, and the text from the Gazetas has the lowest score among the three samples. This is the dimension that is best aligned with our hypothesis, as Semedo reaches an intermediate level of complexity among the three samples. This indicates that, in terms of sentence structure, Semedo has advanced towards a more simplified way of writing. However, it is still not far from Vieira's text.

Table 4 – Syntax complexity analysis (40 metrics).

	Gazetas	Semedo	Vieira
Rank 1 High	10	14	14
Rank 2 Med	6	14	20
Rank 3 Low	24	12	6
Global	**66**	**82**	**88**

Source: created by the authors.

Table 5 reports the specific results of NILC-Metrix for readability scores. In this category, Semedo and Vieira tied for most complex, while the Gazetas reached a lower score. Again, the text from the Gazetas is less complex than both Semedo and Vieira, but, contrary to our hypothesis, Semedo's text is not less complex than Vieira's for this dimension.

Table 5 – Readability complexity analysis (5 metrics).

	Gazetas	Semedo	Vieira
Rank 1 High	0	3	2
Rank 2 Med	3	0	2
Rank 3 Low	2	2	1
Global	**8**	**11**	**11**

Source: created by the authors.

Table 6 contains the results for the psycholinguistic complexity of the texts. This takes into account 20 metrics involving concreteness, familiarity, imageability and age of acquisition of the vocabulary. Similar to what happened with the Readability metrics on table 5, the texts by Semedo and Vieira tied for most complex, and the text from the Gazetas is clearly the least complex in this category. However, because this category is heavily based on vocabulary, it must be considered with caution. Many words that are now considered archaic and less familiar would be more current at that time. However, this should not affect measures of imageability and concreteness.

Table 6 – Psycholinguistic complexity analysis (20 metrics).

	Gazetas	Semedo	Vieira
Rank 1 High	2	8	9
Rank 2 Med	6	8	6
Rank 3 Low	12	4	5
Global	**30**	**44**	**44**

Source: created by the authors.

Finally, Table 7 contains the results for lexical complexity, which was evaluated based on 63 metrics, involving word and word class distributions. The text by Semedo scored the highest here, while the text from the Gazetas is again the least complex among the three samples. Here again, the fact that the vocabulary of the texts is being measured against distributions from modern corpora can have an impact on the results. We can also assume that, by being the only text that consistently uses medical terminology, Semedo's text was penalized here, even if the terms that he used might have been simple by the standards of the period in which the text was written.

Table 7 – Lexical complexity analysis (63 metrics).

	Gazetas	Semedo	Vieira
Rank 1 High	21	22	18
Rank 2 Med	11	28	25
Rank 3 Low	31	13	20
Global	**116**	**135**	**124**

Source: created by the authors.

In general, we can see that the metrics for the text from the Gazetas present a behavior that is compatible our hypothesis, it was considered the least complex text in the sample in all but one of the analyzed dimensions, and even in the Cohesion Complexity dimension, where it scored between Vieira's and Semedo's text, it was still a small gap. In this case, as we already mentioned, the fact that each line of the text usually presents a different piece of news might have penalized its Cohesion score, as the text jumps from one fact to the other without concerning itself with creating a link between them.

Semedo on the other hand ended up being much closer to Vieira's complexity than we first assumed, and it scored the highest complexity both in the general analysis as in several of the individual dimensions. The syntactic dimension was the only one which fits our hypothesis regarding the three samples. This resulted from using simpler and less connectives, combined with shorter sentences than in Vieira's text. However, even in this case the difference between both is not huge, and the text from the Gazetas was clearly the less complex in this category.

Many of the measures used in this paper are based on current Brazilian Portuguese standards, and this needs to be taken into account. For example: the Dale Chall Metric adapted to Portuguese combines the number of unfamiliar words with the average number of words per sentence. The "unfamiliar words" are those that are not included in the basic vocabulary known to today's fourth-year elementary school students in Brazil. For this metric, the entries from a dictionary of simple words by Biderman, Carvalho and Pedroso (2004) were used as reference.

Another similar limitation is represented by the Gunning Fog Readability Index. This measure adds the average number of words per sentence to the percentage of difficult words in the text and multiplies it by 4. The result is directly linked to the 12 levels of American education. This has not been adapted to any Portuguese context. The difficult words, for this metric, are those that have more than two syllables, however, words with more than two syllables are not always difficult in Portuguese (*e.g.*, *árvore* [tree], *professor* [teacher], *escola* [school], *salada* [salad]). This feature is already pointed out as a metric limitation in NILC-Metrix's user guide.

On the other hand, there are adapted measures that we can consider independent of the time or period of production of the texts, such as the Honoré measure. This measure is a kind of type-token-ratio that takes into account, in addition to the number of types and tokens, the amount of *hapax legomena* (*i.e.*, words that occur only once).

Finally, there are a few caveats about texts like Semedo's medical handbook that might influence how complexity could be evaluated. These texts were intended for professional training. We know today that simplifying specialized textual content and terminologies can require more words. As such, for instance, a different evaluation of complexity could be proposed for number of words in the lexical plan, considering that paraphrases and explanations might appear in the text as a simplification strategy. However, on a syntactical level, we can assume that the simplest syntactic order would be the direct form: SUBJECT + VERB + OBJECT + ADVERB. In this sense, considering the tradition of Latin syntax, a less complex sentence is indeed expected: shorter and without subordination. Therefore, whatever would be contrary to this syntactic scheme could receive greater weight in terms of increased textual complexity. These are, however, analyses that would need to be tested empirically, and go beyond the scope of this study.

6 Final remarks

In this paper, we contrasted the complexity in writing of three different texts: one from the *Gazetas Manuscritas* [free translation: *Handwritten News*], one from Semedo's work *Observaçoens medicas doutrinaes de cem casos gravissimos* [free translation: *Medical and doctrinal observations of a hundred severe cases*], and the last one from Fr. Vieira's *Sermons*. We hypothesized that the text from Semedo would score in between the other two in terms of complexity, as it was written with the intention of making scientific works available to people with lower education levels. As we saw in the results, however, Semedo's work ranked most complex in general, and in several of the subcategories, showing that the author was not able to clearly separate his writing from a baroque tradition, which was represented in this paper by Fr. Vieira's work. The work of Semedo only scored in between the other two in terms of syntax complexity, indicating that in this specific linguistic aspect, he was able to improve the

simplicity of his writing, although still far from a more accessible writing, such as the one present in the text from the Gazetas.

Even when available in digitized formats, the understanding of the content of these materials for today's readers is not a trivial task. Thus, by extension, dealing with these materials, transcribed or not, becomes a computational challenge that brings together several other challenges. However, the condition we have, in general, rests on tools adapted to deal with modern language and discourse patterns. In view of this, at first, it becomes important to verify how and how much of these resources, already proven to be efficient in dealing with modern texts, can be used for studies with language from the 17th and 18th century. In this context, it is important to point out that our study did not have the intention to make a deep analysis of the adequacy and eventual need of change of the used metrics to fully characterize the complexity of the 18th century texts. Our main goal was to make a quantitative comparative analysis of the selected texts using, as support, existent natural language processing tools. In this scope, the performed evaluation allowed us to better support a set of conclusions about the comparative complexity of the texts and their authors.

This becomes especially important when we consider the scenario of Diachronic Terminology studies (Dury; Picton, 2009). This scenario is now reinvigorated by the frameworks of Digital Humanities, Digital Philology and the desire to preserve historical heritages, especially the history of science. Santos, Olival and Sequeira (2020) already pointed out a series of multidisciplinary needs in this context.

The automatic representation of lexical and semantic characteristics of old texts, books or corpora will depend on the insertion of a given discourse in a specific socio-historical semantic frame. This process will allow us to understand how and why a given utterance is observed at a given time, as well as in a cultural, ideological and epistemological context. This insertion, which involves a multidisciplinary effort, will imply designing a scenario for the linguistic material under examination, produced at

a given space-time, similar to the comparison that was carried out by Finatto (2020) with a frame for the meaning.

As indicative of this study, by investigating a medical handbook from 1707, we are, in theory, still just at the beginning of the Enlightenment's proposal of facilitating scientific knowledge. Semedo's first book, from 1680 – *Tratado da Peste* [free translation: *Treatise on Plague*][9] –, was already written in the spirit of this ``facilitating ideology''. As such, it would be expected that more than 20 years later, Semedo, being already an experienced and famous physician, would be able to put into practice this idea of facilitating access to knowledge. However, as we see in our data, Semedo's text was still very close to the tradition of a writing style with greater complexity, which is represented by Fr. Vieira's *Sermons*. Still, it should be noted the exception of syntactic complexity, which is lower than Vieira's. The phrasal patterns thus seem to be the point, then, at which he manages to advance towards a more accessible writing for people of lesser erudition.

As future work, one could examine Semedo's last work, the *Atalaya da vida contra as emboscadas da morte* [free translation: *An observatory of life against the traps of death*], published in 1720, when he was already 85 years old. However, this book has a structure similar to that of dictionary entries and does not offer episode reports. Other works by physicians, surgeons and other health practitioners published in the middle or at the end of the century can also be useful for verifying probable features of text simplicity. Another promising source are the nursing manuals from the same period. The first manual of this kind was *Postilla Religiosa, e Arte de Enfermeiros* [free translation: *Religious Postil, and the Art of Nurses*], which was published in 1741 by Fr. Diogo de Santiago[10]. The challenges for these tasks are many and varied. They range from

[9] The whole book is available at: https://books.google.com.br/books/about/Tratado_da_peste.html?id=Z6pIqFOiROwC (accessed on: 22 Jun. 2023).

[10] The whole book is available at: https://archive.org/details/b30507340/page/72/mode/2up (accessed on: 22 Jun. 2023).

preparing the transcribed material, starting from digitized files with or without OCR, to dealing with spelling variation, parsing and text annotation. It is confirmed that it is necessary to have special computational tools and other resources adapted to be able to extract conceptual and linguistic information from texts and corpora from previous centuries.

Another interesting approach would be to make a counterpoint with current texts observed in their greater or lesser complexity. One material that could offer a promising contrast is the present-day text of Brazilian court sentences. This type of discourse tends to be recognized as closely related to the Fr. Antônio Vieira's *Sermons* style, which is very complex in terms of syntactic-semantic constructions, rhetoric features and also terminology (Motta, 2018; Motta, 2021).

It is also in our plans, as future work, to apply and evaluate the proposed methodology to texts from other domains and time periods, aiming to validate its adequacy and generalization capacity.

Acknowledgments

The researchers involved in this study were funded by several international funding bodies. We would like to thank: the Bundesministerium von Bildung und Forschung, for funding project AnGer; the Portuguese Foundation FCT, for funding projects CEECIND/01997/2017 and UIDB/00057/2020; the Brazilian Institutions PPG-LETRAS-UFRGS, for the support, CNPq, for funding a Productivity Research Grant (06/2019, proc. 308926/2019-6) and a research project (Ed. 26/2021 – PDE – Proc. 200051/2023-7), and FAPERGS-CAPES, for funding a research project (06/2018 – INTERNAC., proc.19/2551-0000718-3).

We would also like to thank our colleague Sidney Leal, who is one of the people responsible for the new presentation of the NILC-METRIX system, for helping us in reviewing each of its 200 metrics.

References

ALUÍSIO, S., GASPERIN, C. Fostering digital inclusion and accessibility: the Porsimples project for simplification of Portuguese texts. *In*: **Proceedings of the**

NAACL-HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas, 2010. p. 46–53.

BANZA, A. P., GONÇALVES, M. F. **Roteiro de história da língua portuguesa**. Cátedra UNESCO, Universidade de Évora, 2018, p. 95. Available at: https://core.ac.uk/download/pdf/154812031.pdf. Accessed on: 22 Jun. 2023.

BARBOSA, A. V. Do conhecimento da doença à sua nomeação: uma viagem pelo tratado da conservação da saúde dos povos, de António Ribeiro Sanches. **Panace@**, v. 21(52), p. 37–48, 2020.

BERBER SARDINHA, T.; BARBARA, L. Freqüência e uso de estrangeirismos ingleses no português brasileiro: Um estudo baseado em corpus. **Revista Brasileira de Linguística Aplicada**, v. 5(1), p. 97–114, 2005. DOI https://doi.org/10.1590/S1984-63982005000100006

BIDERMAN, M. T. C., CARVALHO, C. S., PEDROSO, O. **Meu primeiro livro de palavras**: um dicionário ilustrado do português de A a Z. Ática, 2004.

CASELI, H. M., PEREIRA, T. F., SPECIA, L., PARDO, T. A., GASPERIN, C., ALUÍSIO, S. M. Building a brazilian portuguese parallel corpus of original and simplified texts. **Advances in Computational Linguistics, Research in Computer Science**, v. 41, p. 59–70, 2009.

CASTRO, I. **Introdução à história do português**. Edições Colibri, Lisboa, Portugal, 2006.

CUNHA, A. L. V. d. **Coh-Metrix-Dementia**: análise automática de distúrbios de linguagem nas demências utilizando Processamento de Línguas Naturais. 2015. Ph.D. thesis, Universidade de São Paulo, 2015.

DURY, P. ; PICTON, A. Terminologie et diachronie: vers une réconciliation théorique et méthodologique? **Revue française de linguistique appliquée**, v. 14(2), p. 31–41, 2009. DOI https://doi.org/10.3917/rfla.142.0031

FINATTO, M. J. B. Corpus-amostra português do século XVIII: textos antigos de medicina em atividades de ensino e pesquisa. **Domínios de Lingu@gem**, Uberlândia 12(1), 2018. DOI https://doi.org/10.14393/DL33-v12n1a2018-15

FINATTO, M. J. B. Medicina em português no século XVIII: desafios da terminologia diacrônica no cenário das humanidades digitais. **Panace@**, v. 21(52), p. 20–36, 2020.

FINATTO, M. J. B.; QUARESMA, P.; GONÇALVES, M.F. Portuguese corpora of the 18th century: old medicine texts for teaching and research. *In*: **Proceedings of the Conference on Language Technologies and Digital Humanities**. University of Ljubljana, 2018. p. 114–120.

FURTADO, J. F. Tropical empiricism: making medical knowledge in colonial Brazil. *In*: **Science and empire in the Atlantic world**. Routledge, 2008. p. 127–151. DOI https://doi.org/10.4324/9780203933848-8

GAZZOLA, M., LEAL, S. E., ALUISIO, S. M. Predição da complexidade textual de recursos educacionais abertos em português. *In*: **Proceedings of the Symposium in Information and Human Language Technology - STIL**. SBC, 2019.

GRAESSER, A. C.; MCNAMARA, D. S.; LOUWERSE, M. M.; CAI, Z. Coh-metrix: Analysis of text on cohesion and language. **Behavior research methods, instruments, & computers**, v. 36(2), p. 193–202, 2004. DOI https://doi.org/10.3758/BF03195564

LEAL, S. E.; DURAN, M. S.; SCARTON, C. E.; HARTMANN, N. S.; ALUÍSIO, S. M. NILC-Metrix: assessing the complexity of written and spoken language in Brazilian Portuguese. **arXiv preprint**, arXiv:2201.03445, 2021.

LISBOA, J. L.; MIRANDA, T. C.; OLIVAL, F. **As Gazetas Manuscritas da Biblioteca Pública de Évora**. Colibri, CIDEHUS-UE, CHC-UNL, 2002. DOI https://doi.org/10.4000/books.cidehus.3083

LOBENSTEIN-REICHMANN, A. Luther's Contribution as Bible Translator to the German Language. **The Bible Translator**, v. 73(3), p. 301-334, 2022. DOI https://doi.org/10.1177/20516770221140051

MARTINS, T. B.; GHIRALDELO, C. M.; NUNES, M. D. G. V.; OLIVEIRA JUNIOR, O. N. D. **Readability formulas applied to textbooks in Brazilian Portuguese**. 1996. Technical report, ICMSC-USP, 1996.

MOTTA, E. Índices de complexidade textual em sentenças dos juizados especiais cíveis do poder judiciário do estado do Rio Grande do Sul. **Inventário**, v. 1(21), p. 35–50, 2018.

MOTTA, E. Sentenças judiciais e acessibilidade textual e terminológica. **Domínios de Lingu@gem**, v. 15(3), p. 761–813, 2021. DOI https://doi.org/10.14393/DL47-v15n3a2021-6

PIOTROWSKI, M. Natural language processing for historical texts. **Synthesis lectures on human language technologies**, v. 5(2), p. 1–157, 2012. DOI https://doi.org/10.2200/S00436ED1V01Y201207HLT017

QUARESMA, P.; FINATTO, M. J. B. Information extraction from historical texts: a case study. *In*: **Proceedings of the Workshop on Digital Humanities and Natural Language Processing (DHandNLP)**. Co-located with the International Conference on the Computational Processing of Portuguese (PROPOR 2020). Évora, Portugal, 2020. p. 49–56. DOI https://doi.org/10.1007/978-3-030-41505-1

SANTOS, I.; OLIVAL, F.; SEQUEIRA, O. Excavating the data pit: the Portuguese Parish Memories (1758) as a gold standard. *In*: **Proceedings of the Workshop on Digital Humanities and Natural Language Processing (DHandNLP)**. Co-located with the International Conference on the Computational Processing of Portuguese (PROPOR 2020). Évora, Portugal, 2020. p. 69–75.

SANTOS, L. B. D.; DURAN, M. S.; HARTMANN, N. S.; CANDIDO, A.; PAETZOLD, G. H.; ALUISIO, S. M. A lightweight regression method to infer psycholinguistic properties for brazilian portuguese. *In*: **International conference on text, speech, and dialogue**. Springer, 2017. p. 281–289. DOI https://doi.org/10.1007/978-3-319-64206-2_32

SANTOS, R.; PEDRO, G.; LEAL, S.; VALE, O.; PARDO, T.; BONTCHEVA, K.; SCARTON, C. Measuring the impact of readability features in fake news detection. *In*: **Proceedings of the 12th language resources and evaluation conference**, 2020. p. 1404–1413.

SEMEDO, J.C. **Observaçoens medicas doutrinaes de cem casos gravissimos, que em serviço da patria, & das nações estranhas escreve em lingua Portugueza, & Latina Joam Curvo Semmedo**. Officina de Antonio Pedrozo Galram, Lisboa, Portugal, 1707.

SOUSA, M. C. P. d. O Corpus Tycho Brahe: contribuições para as humanidades digitais no Brasil. **Filologia e linguística portuguesa**, v. 16(esp.), p. 53–93, 2014. DOI https://doi.org/10.11606/issn.2176-9419.v16ispep53-93

VERDELHO, T. **Terminologias na língua portuguesa: perspectiva diacrónica**. 1998. Available at: http://clp.dlc.ua.pt/Publicacoes/Terminologias_lingua_portuguesa.pdf. Accessed on: 22 Jun. 2023.

WAGNER FILHO, J. A.; WILKENS, R.; IDIART, M.; VILLAVICENCIO, A. The brWaC corpus: a new open resource for Brazilian Portuguese. *In*: **Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)**. 2018.