HOW TO INTRODUCE STUDENTS TO LINGUISTIC RESEARCH USING VOYANT TOOLS

COMO APRESENTAR A PESQUISA LINGUÍSTICA AOS ALUNOS USANDO O VOYANT TOOLS

Joel Victor Reis Lisboa*
Guilherme Fromm**

ABSTRACT
This paper aims to demonstrate the productivity of using Voyant Tools (SINCLAIR; ROCKWELL, c2023) – a web environment for text analysis – for pedagogical purposes, especially in (future) lexical researchers training. To do so, we propose discussion rounds and practical activities which can be adapted to different target audiences, from basic education to graduate students. We hope this paper encourages new pedagogical uses of Voyant Tools and also guides teachers who want to introduce students to descriptive linguistic research or to use Voyant Tools as one of their pedagogical resources in classroom.
Keywords: Voyant Tools; Corpus Linguistics; Descriptive Linguistics; Researchers training.

RESUMO
Este artigo objetiva demonstrar a produtividade da utilização pedagógica do *Voyant Tools* (SINCLAIR; ROCKWELL, c2023), ambiente *web* de análise textual, na introdução à formação de (futuros) pesquisadores do léxico. Para tanto, apresentamos propostas de rodadas de discussão e de atividades práticas passíveis de adaptação para diferentes públicos-alvo, da educação básica à pós-graduação. Esperamos que este artigo possa instigar novas propostas de utilização do *Voyant Tools* na sala de aula, bem como que sirva de base para professores que desejem introduzir seus alunos nas pesquisas linguísticas de base descritiva ou utilizar o *Voyant Tools* como mais um dos recursos pedagógicos de suas aulas.
Palavras-chave: Voyant Tools; Linguística de *Corpus*; Linguística Descritiva; Formação de pesquisadores.

INTRODUCTION

Corpus Linguistics is a powerful methodology and approach for various fields of linguistic studies, such as Lexicology, Lexicography, Terminology, Terminography, Discourse Analysis, Stylistics, Translation Studies, Language Teaching, Textual Linguistics, Forensic Linguistics, among others (McENERY; WILSON, 2001; O'KEEFFE; McCARTHY, 2010). Its fundamental principles posit that language operates as a probabilistic system, where not all theoretically possible occurrences happen with equal frequency, and that variation across contexts is not random, but rather standardized (BERBER SARDINHA, 2004).

Pedagogical applications of Corpus Linguistics promote a closer link between students and analyses of language as it is authentically produced by speakers in real-life contexts. Moreover, they foster learners' autonomy and greater awareness regarding the probabilistic and associative nature of language, guiding learners towards the recognition and utilization of patterns, ultimately contributing to students' engagement in more authentic linguistic productions (JOHNS, 1991; RÖMER, 2008).

In this paper, we focus on the utilization of Corpus Linguistics, more specifically Voyant Tools[1], a web-based textual analysis environment, in the initial training of (future) researchers in Linguistics. In other words, this paper aims to propose and exemplify pedagogical application possibilities using Voyant Tools, which can be adapted to different target audiences, from basic education to graduate students.

In particular, we suggest three discussion rounds that work as a way of introducing students to lexicon studies, to scientific thinking and to Voyant Tools. We also propose two pedagogical activities to be carried out with Voyant. This paper derives from our experiences as instructors of the workshops "Voyant Tools in the classroom" (LISBOA; FROMM, 2021), held at held at the XI Brazilian School of Computational Linguistics, and

* Doutorando em Estudos Linguísticos pela Universidade Federal de Uberlândia (UFU), Uberlândia, MG, Brasil. joelvictorlisboa@gmail.com. <https://orcid.org/0000-0001-6570-4306>
** Professor Associado do Instituto de Letras e Linguística da Universidade Federal de Uberlândia, (UFU), Uberlândia, MG, Brasil. guifromm@ufu.br. <https://orcid.org/0000-0001-5654-0135>.
1 Available at: https://voyant-tools.org/. Accessed: September 1, 2023.

http://dx.doi.org/10.1590/01031813v63320248644435

"Introduction to Voyant Tools" (LISBOA, 2022), held at held at the II Linguistic Studies Summer School of the Federal University of Uberlândia.

In Section 2, we provide a review of Voyant Tools and 10 tools that can be used in our proposed pedagogical application. In Section 3, we present the theoretical foundations underpinning our proposal. Finally, in Section 4, we outline and discuss the ways in which Voyant Tools can be used in the initial training of (future) researchers in Linguistics.

1. VOYANT TOOLS

Voyant Tools (SINCLAIR; ROCKWELL, c2023) is a free web environment for textual analysis developed by Stéfan Sinclair (McGill University) and Geoffrey Rockwell (University of Alberta). It was launched in 2003 and was originally designed to meet the needs of Digital Humanities scholars (SINCLAIR; ROCKWELL, 2012). However, in practice, it can be used by anyone wishing to perform computer-aided textual analysis.

It enables computer-aided textual analysis, providing, in a few seconds, linguistic and statistical information about the corpora under examination. Its 28 tools allow multidimensional analysis and make it possible to retrieve multiple information at different levels of corpora. Thus, it is a useful resource for macroanalysis and targeted microanalysis, which are routine procedures in Corpus Linguistics research.

The user-friendly and striking layout, the interactivity among its tools and the ease of sharing data are some of the key points that set Voyant Tools apart from other widely used linguistic analysis software, such as WordSmith Tools (SCOTT, 2020) and AntConc (ANTHONY, 2020).

Considering that Voyant Tools represents a combination of intuitiveness and several configurable analytical tools capable of providing information at different levels of the corpora, it is useful for a variety of target audiences, such as in-training researchers, teachers, academics and students, as well as beginners and advanced users (WELSH, 2014; ALHUDITHI, 2021). As such, Voyant is a great introductory resource to Corpus Linguistics and Lexicon Studies.

1.1 Advantages and limitations

Previous research that took Voyant Tools as study object or which used it as an analytical resource has evidenced some of its highlights, such as the following:

- Cost-free and no download[2] or login required (SAMPSEL, 2018; ALHUDITHI, 2021; SÁNCHEZ TARRAGÓ, 2021).
- Several file formats accepted for processing (SÁNCHEZ TARRAGÓ, 2021), such as .pdf, .txt, .doc, .docx, .xls, .xlsx, html, .xml, .odt, .rtf, .pages, as well as zipped files.
- User-friendly layout, which does not require high instrumental or computational training, and appealing data visualization formats (WELSH, 2014; SAMPSEL, 2018; ALHUDITHI, 2021; HETENYI; LENGYEL; SZILASI, 2019; MILLER, 2018; HENDRIGAN, 2019).
- Interactivity among its tools (SAMPSEL, 2018; MILLER, 2018; SINCLAIR; ROCKWELL, 2012), so there is no need to switch pages and windows like in other lexical analysis software.
- Possibility of creating URLs for the tools (with results included) and for whole projects (WELSH, 2014; SAMPSEL, 2018; MILLER, 2018). This means that, by creating URLs, there is no need to re-upload corpora or rearrange/reconfigure tools. In addition, this also favors remote joint work.
- Files and data exportation in various formats (WELSH, 2014), such as .txt, .svg, .xml, in the original format of the uploaded files, as static images (.png), as HTML snippets for interactive visualization[3] etc.

2 However, there is the Voyant Server, a standalone version for desktop which is available at: https://github.com/sgsinclair/VoyantServer#voyant-server. Accessed: September 1, 2023.

3 An example of embedding interactive visualization on websites is available at: https://blogs.reed.edu/ed-tech/2017/03/text-analysis-using-voyant-tools/. Accessed: September 1, 2023.

- Extensive documentation (WELSH, 2014; ALHUDITHI, 2021). Besides that, its manual[4] has interactive visualizations, which make explanations much more productive as they allow users to browse the tools in the manual's environment.
- Flexibility of use on different devices (computers, tablets and smartphones) and operational systems (Android, iOS, Mac OS, Linux e Windows) (ALHUDITHI, 2021).

Furthermore, the web environment interface is available in 14 languages[5] and, theoretically, Voyant Tools can process texts in any language due to its tokenization algorithms.

Evidently, like any other textual analysis software, Voyant Tools has some limitations. In addition to the manual being available only in English, which restricts the number of users who can use it, previous research has shown the following limitations:

- Slow loading of texts/files (WELSH, 2014; SAMPSEL, 2018). According to tests we performed, we concluded that slowness may be related to the quality of the equipment/connection, file format or corpus extension.
- Complexity in obtaining information in some tools, such as Knots and Mandala (WELSH, 2014), which have more complex results presentation formats.
- The terminology used by Voyant Tools can be a handicap for beginners (ALHUDITHI, 2021). Besides that, Voyant makes use of some well-stablished terms in areas such as Phraseology and Terminology (for instance, "collocates" and "terms"), but linked to different concepts from those of these areas, which can confuse some users. However, although the manual does not have a glossary, the information about each tool is very detailed and minimizes the possibility of confusion.
- Analyzing confidential data can be high risk, as there is not much information available concerning data protection and storage by the system (ALHUDITHI, 2021). However, there are two possibilities for reducing information leakage risk: (i) creating a password to access the corpus or (ii) using Voyant Server, the standalone desktop version.
- Lack of a greater variety of word lists (ALHUDITHI, 2021). Currently, Voyant Tools has three indexed corpora (Austen's Novels, Shelley's Frankenstein and Shakespeare's Plays) and they are the ones used as reference corpora for the English language. The addition of a greater variety of word lists "would generate meaningful comparisons across different texts and accommodate a [greater] range of research purposes" (ALHUDITHI, 2021, p. 49).
- Lack of finer details on its metrics (OLIVEIRA; BRITO; OLIVEIRA, 2018). Although the manual is extensively detailed to some extent, it really lacks more in-depth information about Voyant's statistical metrics.

In addition, the manual does not detail which reference corpora are used to extract keywords for languages other than English. Moreover, Voyant provides stoplists for 35 languages, but there are no details regarding the criteria adopted for their creation.

1.2 The tools

Since its launch in 2003, Voyant Tools has gone through several redesigns and enhancements, where tools have been discontinued, some have been introduced, and others have been improved. Currently, Voyant Tools has 28 analytical tools. In the following section, we briefly detail the 10 most productive for the pedagogical activities presented in this paper.

1.2.1 Summary

The Summary tool gives users information about the corpus and the files. It is divided into six sections, as shown in Figure 1.

4 Available at: https://voyant-tools.org/docs/#!/guide. Accessed: September 1, 2023.
5 Namely: Arabic, Bosnian, Croatian, Czech, English, French, German, Hebrew, Italian, Japanese, Portuguese, Russian, Serbian and Spanish.

Figure 1. Corpus Teen Wolf[6] results in the Summary tool
Source: Voyant Tools.

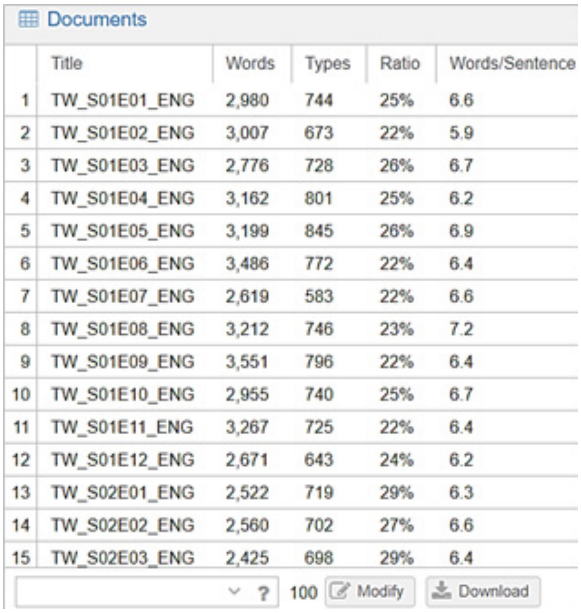
The first section describes the total number of files, tokens[7] and types[8], as well as the corpus creation date, whereas the second shows the files with the largest and smallest extensions, along with their number of tokens. The files with the higher and lower vocabulary density as well as their type/token ratios[9] are presented in the third section. The fourth section exhibits the files whose average words per sentence is higher and lower, along with the average value of each. In the fifth section, files whose readability indexes are higher and lower are listed, that is, files that are theoretically easier or more difficult to read when considering the word length metrics and the average number of words per sentence. The most frequent words in the corpus and their total occurrence are presented in the sixth section. The seventh and last section lists the words with the highest keyness value[10] in each file, along with their frequency.

This tool allows users to change the number of elements to be displayed (min. 5 and max. 59). However, the seventh section list is limited to the first 20 corpus files.

1.2.2 Documents

The Documents tool gives information about the corpus constitution, details of its files, as well as functionalities to modify it. It is displayed in the form of a table with five main columns, as shown in Figure 2.

6 We compiled this corpus for the aforementioned workshops, with the aim of demonstrating the possibilities of using Voyant tools for pedagogical purposes. This corpus consists of English subtitle files from the 6 seasons of the Teen Wolf TV series.
7 All items or words, including repetitions.
8 Different items or words, that is, disregarding repetitions.
9 Briefly, type/token ratio – obtained from the division of the number of types by the number of tokens – is used to calculate the lexical diversity of a corpus. The higher the ratio, the more varied is the vocabulary in the corpus.
10 Briefly, keyness is an index obtained through statistical tests (log-likelihood or chi-square) carried out with two word lists (study and reference corpus). It indicates words whose relative frequency is statistically significant in the study corpus (BERBER SARDINHA, 2004).



Documents

	Title	Words	Types	Ratio	Words/Sentence
1	TW_S01E01_ENG	2,980	744	25%	6.6
2	TW_S01E02_ENG	3,007	673	22%	5.9
3	TW_S01E03_ENG	2,776	728	26%	6.7
4	TW_S01E04_ENG	3,162	801	25%	6.2
5	TW_S01E05_ENG	3,199	845	26%	6.9
6	TW_S01E06_ENG	3,486	772	22%	6.4
7	TW_S01E07_ENG	2,619	583	22%	6.6
8	TW_S01E08_ENG	3,212	746	23%	7.2
9	TW_S01E09_ENG	3,551	796	22%	6.4
10	TW_S01E10_ENG	2,955	740	25%	6.7
11	TW_S01E11_ENG	3,267	725	22%	6.4
12	TW_S01E12_ENG	2,671	643	24%	6.2
13	TW_S02E01_ENG	2,522	719	29%	6.3
14	TW_S02E02_ENG	2,560	702	27%	6.6
15	TW_S02E03_ENG	2,425	698	29%	6.4

? 100 Modify Download

Figure 2. Corpus Teen Wolf results in the Documents tool
Source: Voyant Tools.

The five main columns present, respectively: (i) file name; (ii) number of tokens; (iii) number of types; (iv) type/token ratio; (v) average number of tokens per sentence. Users can activate more columns which exhibits files metadata[11], such as: authorship, date, publisher, place of publication, keyword, collection and language.

This tool also allows users to make changes to the corpus, such as reordering, deleting or adding files, as well as creating a new corpus from specific files. In addition, it makes it possible to download the corpus in the original upload format, .txt or .xml, which facilitates sharing corpora with collaborators. Two other available features are filtering files based on search words and reorganizing columns in ascending or descending order.

1.2.3 Reader

The Reader tool enables the complete reading of the corpus files, presenting information about the frequency and word distribution in the files and throughout the corpus as a whole. As shown in Figure 3, when hovering the mouse over each word, a pop-up appears indicating the raw frequency of the word in the file under analysis.

11 Metadata is inserted when the corpus is created. Therefore, if it was not inserted in the corpora creation environment, it will not be displayed in this tool.

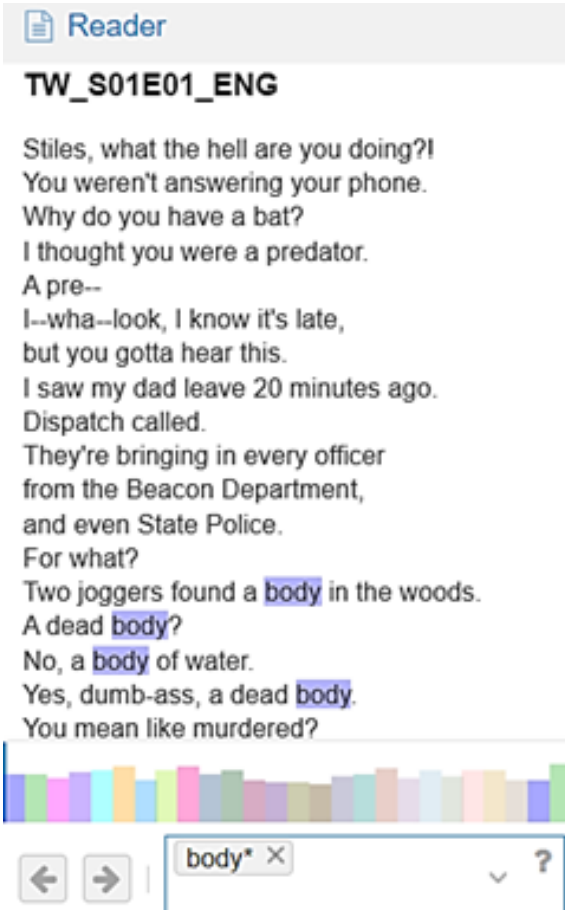


Figure 3. Corpus Teen Wolf results in the Reader tool
Source: Voyant Tools.

This tool allows users to search for specific words, which are marked in blue. Furthermore, at the bottom of the tool there are blocks that represent each of the files. The larger and wider the block, the larger the file.

One of the most productive aspects of this tool is that it responds to the use of other tools, such as Summary, Documents, Contexts, Cirrus, Links, Trends etc. Hence, by clicking on a specific word or file in other tools, the Reader automatically updates, directing users to read the file, while highlighting the occurrences of the selected word or showing its distribution in the corpus through a sparkline graph[12].

1.2.4 Contexts

The Contexts tool works as a concordancer, that is, it lists all occurrences of a given word in their linguistic cotexts[13]. The data visualization is in the form of a four-column table, as shown in Figure 4.

12 Sparkline is a type of graph used to visually represent trends and variations in values.
13 The immediate linguistic environment that surrounds a given word to the left and right.

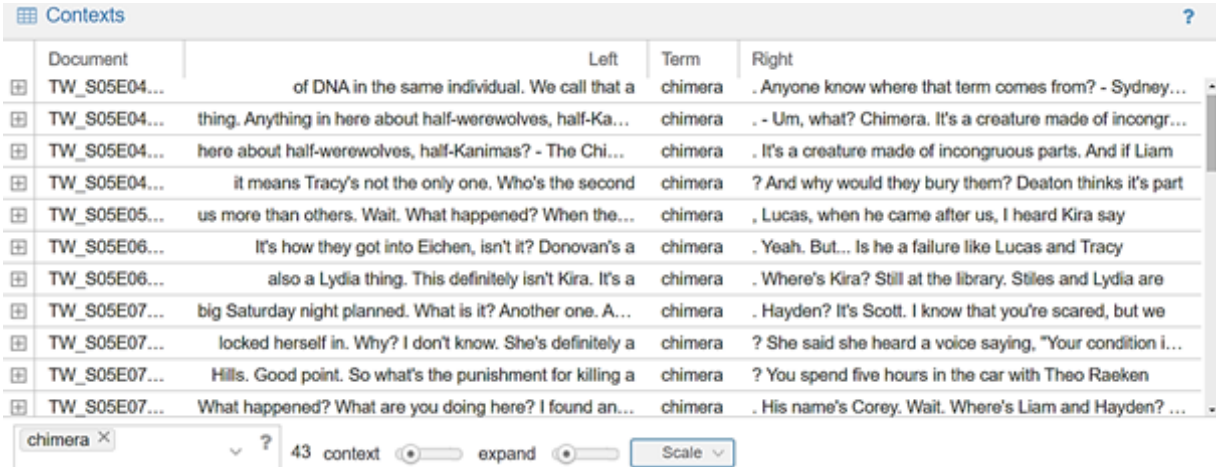


Figure 4. Corpus Teen Wolf results in the Contexts tool
Source: Voyant Tools.

The columns indicate, respectively: (i) file name; (ii) items to the left; (iii) search node; (iv) items to the right. Moreover, users can activate a fifth column which indicates the search node position (based on each file token index). The tool also makes it possible to expand concordance lines to examine the linguistic context in which each search node occurs.

Some resources available are: (i) search for specific word(s); (ii) dimensioning of results based on the entire corpus or specific files; (iii) cotext extension adjustment[14] (min. 5 and max. 50 tokens); (iv) expanded context length adjustment (min. 5 and max. 500 tokens) and (v) reorganization of columns in ascending or descending order.

1.2.5 Cirrus

The Cirrus tool displays a word cloud based on the frequency of occurrence of words in a corpus or in specific files. The larger the word in the cloud, the more frequent it is compared to the other words. The most frequent words are centralized, however, the absolute position is irrelevant. Figure 5 illustrates how this tool works.

Figure 5. Corpus Teen Wolf results in the Cirrus tool
Source: Voyant Tools.

By hovering the mouse over each word, a pop-up appears informing its number of occurrences. In addition, the tool has the following features: (i) determination of the number of words in the cloud (min. 25 and max. 500

14 For a better visualization of the results, the ideal number of tokens on the left and right is between 10 and 15.

words); (ii) cloud scaling based on the entire corpus or specific files; (iii) application of stoplist[15] and/or whitelist[16]; (iv) colors and fonts specification, and (v) creation of word categories.

1.2.6 Links

The Links tools displays a network graph that represents links between words that tend to occur with high frequency. This tool is illustrated in Figure 6.

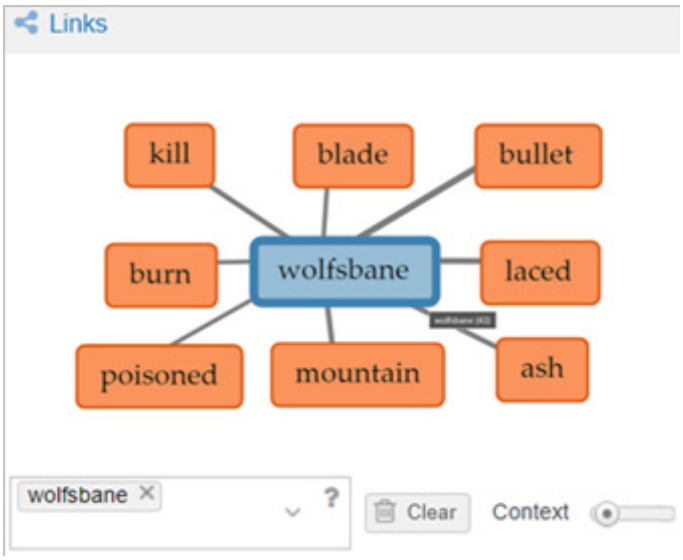


Figure 6. Corpus Teen Wolf results in the Links tool
Source: Voyant Tools.

This graph is made up of rectangles that represent the base words (blue) and their collocates (orange). The link between them indicates that the words tend to co-occur frequently. By hovering the mouse over a base word, the tool informs, through a pop-up, its absolute frequency. When positioning the mouse over a collocate, the tool shows the number of times it co-occurs with the base word, considering the cotext extension specified by the user.

Other features available in this tool are: (i) search for specific word(s); (ii) cotext extension adjustments (min. 3 and max. 30 tokens); (iii) repositioning of the graph rectangles and (iv) stoplist application.

1.2.7 Collocates

The Collocates tool presents a table with words that tend to co-occur and their co-occurrence frequency. By default, the table is made up of three columns that display, respectively, the base word, the collocate and the number of co-occurrences, considering the cotext extension specified by the user. This tool is shown in Figure 7.

15 Stoplist is a list of words to be disregarded by the tools. We suggest caution when using stoplists provided by Voyant Tools, as their creation principles are not specified in the manual. Furthermore, we noticed some inconsistencies in the Portuguese and English stoplists.

16 A whitelist is the opposite of a stoplist, as it consists of a list of words to be considered by the tool. That is, only the data of the words in this list is to be presented by the tool.

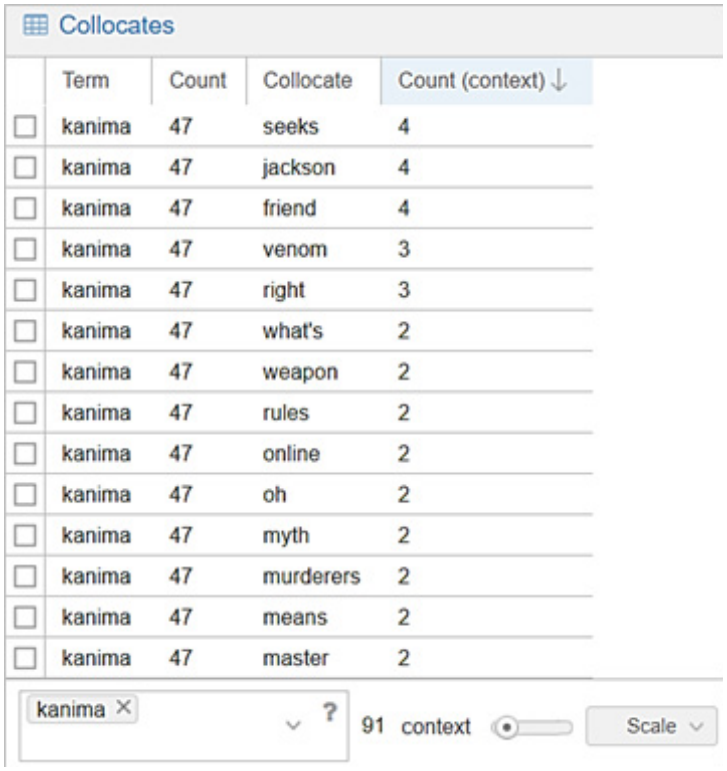


Figure 7. Corpus Teen Wolf results in the Collocates tool
Source: Voyant Tools.

Users can activate a fourth column (positioned after the base word column, as shown in Figure 7) which presents the number of occurrences of the base word, enabling a pre-analysis of the mutual attraction tendency between the base word and its collocate. In addition, the tool has the following features: (i) search for specific base-word(s); (ii) dimensioning from the entire corpus or from specific files; (iii) cotext length adjustment (min. 1 and max. 30 tokens between the base and the collocate) and (iv) columns reorganization in ascending or descending order.

1.2.8 MicroSearch

The MicroSearch tool informs the distribution of the occurrence of certain words throughout the corpus files. This tool is illustrated in Figure 8:

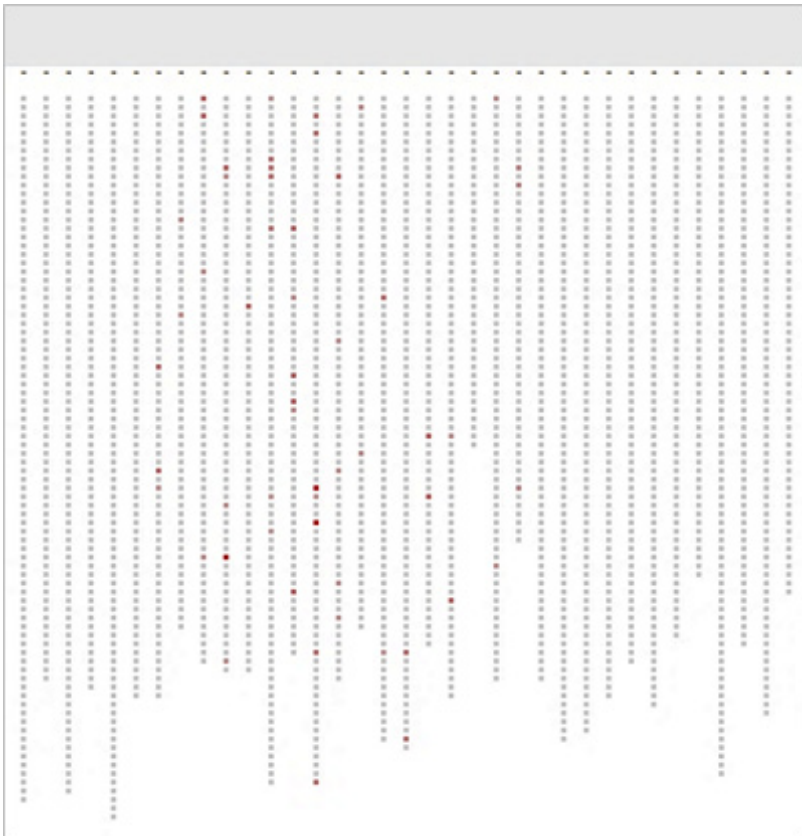


Figure 8. Corpus Teen Wolf results in the MicroSearch tool
Source: Voyant Tools.

Each file corresponds to a column. The longer the column, the longer the file size. Each search word is represented by a red square, which is positioned where the word occurred in each file. Therefore, this tool makes it possible to analyze the distribution of each search word throughout files (if it is more prominent at the beginning, middle or end), as well as to verify which files (do not) contain the specified word(s).

1.2.9 Bubblelines

The Bubblelines tool indicates the frequency and distribution of words in a corpus or in specific files. Each file corresponds to a horizontal line. They are also subdivided into identical segments based on the number of tokens. When entering search words, these words are represented by bubbles. These bubbles are distributed along the horizontal line, symbolizing words distribution along the file. The bigger the bubble, the higher the word frequency in the specific segment. In addition, the tool makes it possible to check both the frequency in specific segments and the total frequency of each search word in each file. Figure 9 illustrates this tool.

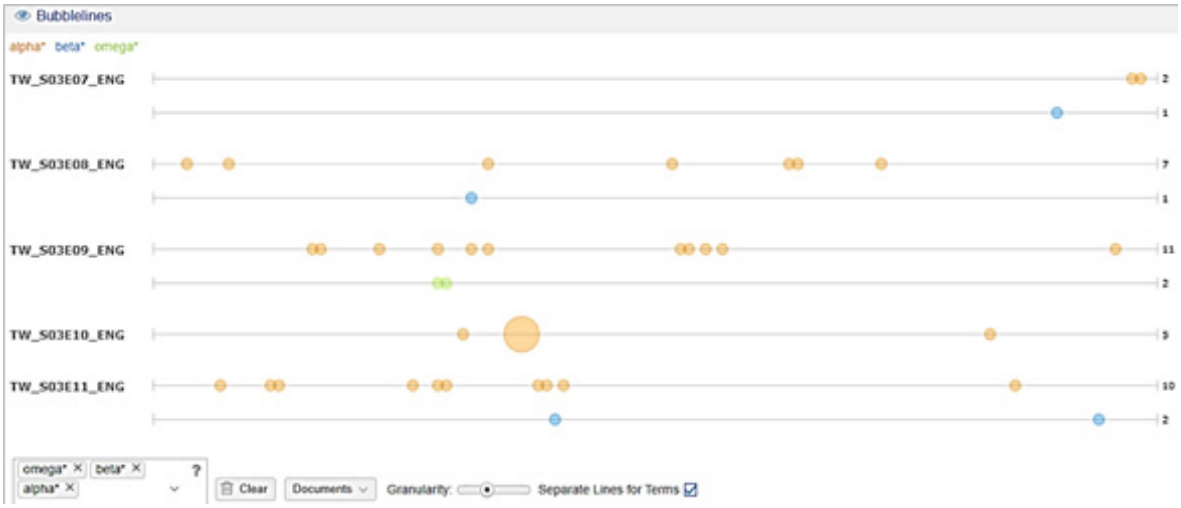


Figure 9. Corpus Teen Wolf results in the Bubblelines tool
Source: Voyant Tools.

Its available resources are: (i) search for specific word(s); (ii) results dimensioning based on the entire corpus or specific files; (iii) granularity adjustments, that is, the number of identical segments that the corpus/files are to be subdivided (min. 10 and max. 300 segments); (iv) search words visualization in the same or separate lines; (v) color palette adjustment and (vi) creation of word categories.

1.2.10 Trends

The Trends tool displays a graph showing the distribution of word occurrences throughout the corpus, in which the Y axis exhibits frequency information (absolute or relative frequency) and the X axis presents the files names, as shown in Figure 10.

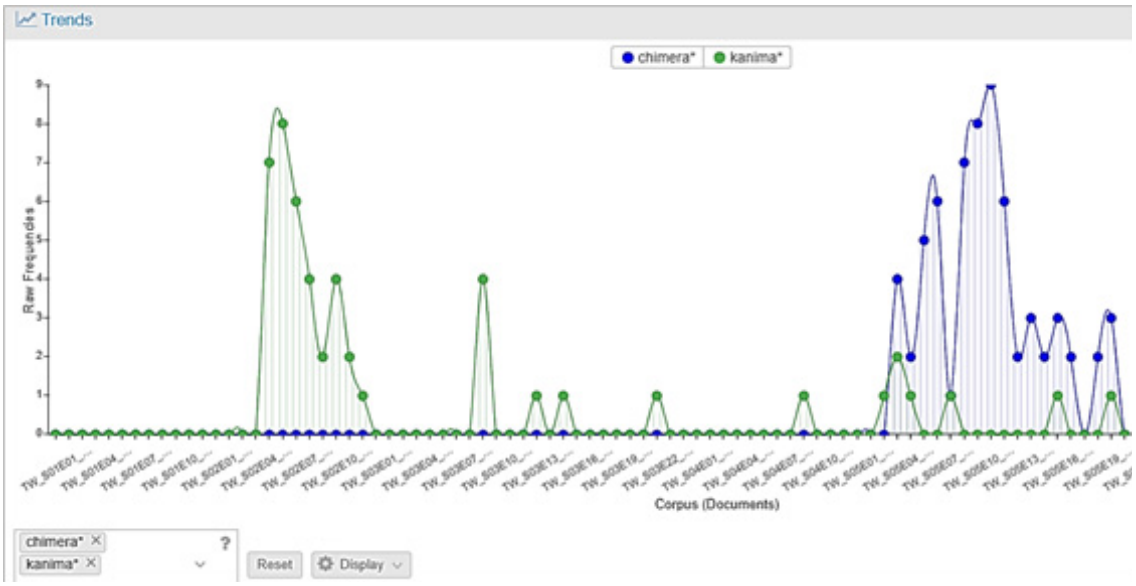


Figure 10. Corpus Teen Wolf results in the Trends tool
Source: Voyant Tools.

By hovering the mouse over certain points of the graph, pop-ups appear indicating the word, the number of occurrences and the files names (or segment number, in cases in which the corpus consists of a single file). Its additional resources are: (i) search for specific words; (ii) chart type selection (area, columns, line, stacked bars or

stacked bars + lines); (iii) adjustment of the number of segments (min. 2 and max. 100[17]); (iv) alternation between absolute and relative frequency; (v) color palette adjustment and (vi) creation of word categories.

2. NOTES ON LINGUISTIC RESEARCH TRAINING

In this section, we briefly make some considerations about the initial training of researchers, on which our pedagogical proposals with Voyant Tools are based. Our discussion is organized around four main points: (i) adaptation as a fundamental aspect; (ii) student protagonism; (iii) methodical thinking and documentation; (iv) training and awareness-raising.

2.1 Adaptation as a fundamental aspect

Depending on the target audience and the students' educational background, using a ready-made corpus and focusing on more practical processes of how to use the tools and their resources for linguistic analysis would be ideal. This is because, as noted by Sinclair and Rockwell (2012), by going straight to corpora analyses, we can skip more complex procedural parts about corpora compilation and treatment. The focus on these highly technical notions and pre-analysis procedures, depending on the target audience, may hinder the pedagogical process. Furthermore, due to gaps in the computational literacy of some students, the use of technology to perform linguistic analyzes can be something, *a priori*, frightening and highly complex. So, what are the advantages of making this learning process even more complex?

Evidently, this is not entirely valid when it comes to undergraduate research scholars and graduate students, who are expected to master all the procedures of scientific research, from hypotheses formulation to the actual analyses, description and data discussion (LORCH, 2016). As Perini (2006) observes, there are limits to simplification, and one should not lose sight of the formative character in the introduction to linguistic research. On the other hand, these considerations are important to bear in mind when dealing with elementary and high school students, as well as university students at the beginning of undergraduate education.

2.2 Student protagonism

Ideally, the teacher should guide the students in the process of formulating their research questions/problems and in the systematization of the methodological processes to be performed, in order to encourage critical thinking and student autonomy. The research problem students come to formulate will trigger hypotheses formulation, decisions on what would be the ideal study corpus, the search for more relevant and productive tools and resources for a specific research, adjustment and reformulation of hypotheses etc. (PERINI, 2006; FERNÁNDEZ PÉREZ, 2011). After all, these are assertive decisions expected of a researcher.

When dealing with research training of undergraduates, undergraduate research scholars and graduate students, teaching pre-analysis technical procedures is essential, so that they can compile their own study corpora. Furthermore, as Sinclair and Rockwell (2012, p. 250) point out, "students should be encouraged to have a set of questions and hypotheses before they even touch the computer, since this forces them to look for the tools that might help them answer their questions".

2.3 Methodical thinking and documentation

Sinclair and Rockwell (2012) emphasize the importance of teaching students to think methodically and to carefully document all steps taken in the research/project. According to the authors, the interactivity of Voyant Tools, while being a facilitator, can become a handicap in the process of writing a research report. Due to the numerous possible ways to get to certain results, students may struggle in satisfactorily proving their arguments or in demonstrating how they arrived at certain conclusions.

Furthermore, teaching students to think methodically and to document the research step by step is to train them for replicability (a priceless principle for scientific research) and to understand the *modus operandi* of the

17 This feature is only used if the corpus contains a single file.

tools they use, so that they are not passive users without minimal understanding of how their computational resources work.

2.4 Training and awareness-raising

Teaching students to use resources like Voyant Tools is relevant, especially with regard to training for linguistic research. But our duty as trainers of (future) researchers is not only that. As Mahony and Pierazzo (2012, p. 224) state, "skills training is not research training", so they need to be trained for new ways of conceiving and seeing the world, for learning thoroughly "new methodologies and new ways of thinking".

In the case of training (future) (applied) linguists, we emphasize language awareness and decision-making training with regard to linguistic research. This does not only require instrumental skills, but critical and reflective thinking about language and research processes (PERINI, 2006; SHEPHERD, 2012; LORCH, 2016).

3. VOYANT TOOLS IN THE TRAINING OF (FUTURE) RESEARCHERS

Prior to the effective use of Voyant Tools in the classroom by students, it is relevant that some concepts and assumptions are evidenced and clarified. To this end, the ideal is that, in previous classes, discussion rounds are held so that these concepts and assumptions are evidenced and understood. We present below three discussion rounds ranging from more general to more specific discussions. We stress that they can be adapted according to the specialization level of the target audience, despite being equally relevant for research training at different educational levels.

3.1 Discussion Round 1: micro and macroanalysis

For Round 1, we propose the following topics and questions:

- Brief introduction to the concepts of micro and macroanalysis.
- How does macroanalysis differ from microanalysis in practice?
- What is the usefulness of macroanalysis? What does it fail to cover?
- What are the (dis)advantages of analyzing a corpus compared to analyzing a single text?
- Would it be possible to combine micro and macroanalysis?

The topics and questions raised in Round 1 aim to promote reflection on alternative ways of reading beyond the conventional one, focusing on the advantages, disadvantages and differences between micro and macroanalyses. We here use the terms "microanalysis" and "macroanalysis" according to Jockers (2013), although they come from what Moretti (2005) has named "close reading" and "distant reading".

In short, microanalysis (or close reading) refers to the way in which we normally read and analyze texts on paper or computer. It is a horizontal and more detailed reading, which focuses on one text at a time and in which the development of events, ideas and arguments are observed in a deeper and more individualized way (JÄNICKE et al., 2015). This type of reading also involves skimming and scanning techniques and detailed analysis of figures, charts, tables etc.

On the other hand, macroanalysis (or distant reading) consists of a more vertical and panoramic computer-aided reading of groups of texts. It usually involves identifying linguistic items association frequency, observing corpora idiosyncratic features and identifying patterns whose recurrences are difficult to be identified and/or quantified manually.

Round 1 also leads students to an extension of these two concepts, which is targeted microanalysis, a combination of the two previous approaches. Microanalysis is time-consuming and generally makes it difficult to identify and quantify more panoramic idiosyncratic patterns. Macroanalysis, due to its more panoramic nature, generally disregards less recurrent but highly relevant traits, which can generate a distorted image of the corpus under analysis. Since macroanalysis expands the possibilities and limits of microanalysis and because microanalysis allows better contextualization of results and focus on aspects that are not so recurrent, but

fundamental, the combination of the two approaches is a powerful process (JOCKERS, 2011), and Voyant Tools enables this multiple analysis.

Therefore, when we make use of computational resources to analyze corpora, we can perform targeted microanalyses, going directly to the most relevant information for a given research objective. And it is from the macroanalysis carried out using the computer that we identify those focal points. It is basically this process that research in Corpus Linguistics, especially corpora-driven research, performs.

It is relevant that students begin to reflect on these issues prior to the effective use of Voyant Tools. After this more general discussion on different types of reading and analyses, that aims to prepare students to use Voyant Tools, we can move on to the next discussion round.

3.2 Discussion Round 2: on words and frequencies

Round 2 addresses occurrence frequency, words distribution and idiosyncratic characteristics of some corpora. For this discussion round, we suggest the following topics:.

- What are prominent words?
- What words are expected to be prominent in a corpus? Why?
- How would these words be distributed in the texts?
- Are all prominent words relevant?
- Introduction to stoplists. What is the use of a stoplist?
- What do the most frequent words allow us to infer about a corpus or a text?

In this discussion round, the idea is to instigated students to anticipate what can be found in given corpora, as well as to realize the relevance of frequency for descriptive studies. In addition, this round also aims to stimulate reflections on what would be relevant and what, although frequent, would be of minor interest depending on the research objectives.

Round 2 can be carried out using Voyant Tools, so that practical examples are given. As an example, Round 2 can follow a specific theme, such as cooking. In this case, the questions must be adapted: “What words are expected to be prominent in a corpus of recipes? And in a corpus of dessert recipes? How would these words be distributed throughout the texts?”

In order to provide practical examples, the teacher must have a corpus of recipes previously compiled, treated and processed by Voyant Tools. As the students contribute to the discussion, through their answers to the questions, the teacher can display the results in Voyant Tools itself. This process of anticipating the linguistic data that can be found in a corpus, as well as formulating hypotheses about the corpora, is relevant with regard to training for language research and description.

Another aspect addressed in Round 2 is the differentiation between central and peripheral data. A more practical proposal, already using Voyant Tools, would be to display a list of words from the corpus of recipes, but without applying a stoplist. This listing can be obtained from the Summary tool. Students will notice that the most frequent words are the grammatical ones. The teacher can guide the discussion asking if these words from the list could help in the description of the vocabulary of culinary recipes and what would be the way of excluding grammatical words so that we can focus only on the lexical ones. These questions would already introduce the other discussion topic, the use of stoplists. By doing so, the teacher would now be able to exemplify in practice how and why a stoplist is used. One discussion ends up leading to the other.

Still using the corpus of the given example, it is possible to approach the distribution of words in the texts using the Trends, MicroSearch or Bubblelines tools. This discussion can be initiated by asking which segments of the recipes would have the highest recurrence of nouns, verbs or adverbs. Students would formulate hypotheses and justify their answers. Then, the teacher would present the data in one or some of the mentioned tools.

Regarding the last question (What do the most frequent words allow us to infer about a corpus or a text?), again, Voyant Tools itself can be used. The teacher can take a corpus already compiled, treated and processed by Voyant. Students, without knowing the origin of the corpus, must formulate hypotheses about its content from the results in the Cirrus tool, for example, with stoplist activated. This is a productive exercise, as students are led to formulate hypotheses based on linguistic data, a common procedure of corpora-driven research.

In Round 2, in addition to the discussions raised, students will have had the opportunity to see the tools in actual use and hypothesize based on the results presented. In the next discussion round, the idea is that students are led to understand how these tools work to show the results seen throughout Round 2.

3.3 Discussion Round 3: the *modus operandi*

Round 3 aims to raise awareness of the processes made by the tools to present the results. This understanding is essential to stimulate methodical thinking and active attitudes towards the tools that they will use in the pedagogical application proposals detailed in the next section. The focal points of Round 3, which can be adapted and extended, are the following:

- What is the process that the X tool does to present the results?
- How did you arrive at this conclusion?
- How can we use the results of X, Y and Z tools to prove our arguments?
- Brief introduction to charts and tables reading.

All of the tools detailed in Section 2.2, and the other 16 tools that make up Voyant Tools, can be used in this discussion round. Evidently, it is important that the teacher knows well the operation and resources available in each tool to be addressed, as we also described in Section 2.2. The teacher can also count on the extensive and detailed Voyant Tools manual[18].

3.4 Activity Proposal 1: formulating data-based hypotheses

This first proposal was designed to a more initial level, for elementary and high school students, but it can also be expanded and adapted for undergraduate research scholars, undergraduates and graduate students. Its step by step is as follows:

(i) Discussion Rounds 1 and 2.
(ii) Brief introduction to *Voyant Tools*.
(iii) Sharing a link with an uploaded corpus and a specific skin[19] configured.
(iv) Students explore the tools/results and formulate hypotheses about the content of the corpus, and their arguments must be data-oriented.

The discussions on micro and macroanalysis (Round 1), as well as on words and frequencies (Round 2), are essential to this activity proposal, along with a brief introduction to Voyant Tools. This way, students are prepared beforehand.

The corpus to be used can be of different nature: it can be a corpus of song lyrics from a specific musical genre, makeup blogs, a specialized corpus, transcripts of sports championships, e-sports etc. The important thing is that students do not know its content and that they formulate data-based hypotheses, as well as defend their arguments also based on the data. In addition, we recommend that students work in groups, as this is a great opportunity to stimulate collaborative skills.

The most productive tools for this activity are Summary, Documents, Contexts, Cirrus and Links. We suggest that the stoplist resource be activated in the tools, as it will facilitate the analysis made by the students. Furthermore, we emphasize that the Reader tool should not be used, given that it allows the full reading of the files that make up the corpus, thus not meeting the aims of this activity.

It is worth noting that Voyant Tools runs on different devices and operational systems. Therefore, students do not necessarily need to be all using computers. Furthermore, internet access in the classroom is not a crucial factor, as Voyant Tools can be used in its standalone version. In addition, the teacher can also take screen shots of the tools results to the classroom in case there is no internet access. Another point worth mentioning is that this

18 Available at: https://voyant-tools.org/docs/#!/guide. Accessed: September 1, 2023.
19 Skin is a term used by Voyant Tools referring to the combination of tools displayed in its interface. The default skin comprises five tools (Cirrus, Reader, Trends, Summary and Contexts), but, according to the user's needs, it is possible to combine different tools to create new skins.

activity proposal can be carried out both in classroom and at home, given that Voyant allows users to share links with the uploaded corpus and a pre-configured skin.

3.5 Activity Proposal 2: writing a descriptive report

Although this proposal can be adapted for elementary and high school students, it was originally designed for undergraduates and graduate students. In addition, it is based on what Fromm (2020) conceives as Pedagogical Terminography. Its step by step is the following:

(i) Introduction to Corpus Linguistics.
(ii) Brief introduction to Voyant Tools.
(iii) Discussion rounds 1, 2 e 3.
(iv) In pairs, students compile a subtitle corpus of science fiction/fantasy TV series.
(v) Students select representative terms from the fictional world under examination.
(vi) As a final task, students write a descriptive report both of the corpus and of the selected terms.

As this is a more complex activity, in addition to the discussion rounds and a brief introduction to Voyant Tools, an introduction to Corpus Linguistics is also necessary. After these introductory discussions, students will be better prepared to begin their projects.

Science fiction and, mainly, fantasy TV series have rich and highly neological lexicon, which is particularly productive for descriptive linguistic studies. In addition, they generally enjoy widespread popularity, especially considering the growing number of streaming platforms. These factors indicate the usefulness of working with subtitle corpora of these genres, as they aggregate lexical richness, neological creations and popularity among students.

One potential downside of working with subtitle corpora in Voyant Tools is that it does not ignore subtitle timestamps, thus hindering the analysis. However, considering the target audience of this proposal, this is an opportunity for students to put into practice the criteria and methods of corpora cleaning, which are addressed in the Corpus Linguistics introductory step. For other target audiences, this hurdle can be overcome by using Aspose's free online .srt to .txt converter[20], since timestamps are already automatically removed in the conversion process.

Regarding the terms to be selected for analysis, they should preferably be characteristic and fundamental to the fictional universe of the chosen TV series. They may even coincide with words from the general language, but they must convey meanings that allow a greater understanding of the fictional universe under analysis. For example, within the scope of the Terminology in Fiction project, created and coordinated by Prof. Dr. Guilherme Fromm at the Federal University of Uberlândia, among the terms selected by graduate students who analyzed the subtitles of 3% (a Netflix production) are "cause", "inland" and "shell", lexical items of the general language, but which in the fictional universe of the series acquire meanings different from those found in general language dictionaries[21].

Students should be encouraged to document all procedures performed throughout the project, due to what we mentioned in Section 3.3 of this article. In the descriptive report, students must present the following data: (i) corpus description: genre, number of files, extension (in types and tokens); (ii) terms description: absolute and relative frequency, files in which they occur the most, distribution in the corpus, co-occurrences, as well as corpus-driven definitions. All tools detailed in Section 2.2 can be used in this activity.

Finally, and once again, we emphasize the compatibility of Voyant Tools with different devices and operational systems, the possibility of the project not necessarily being carried out in class (due to the ease of sharing links to projects and skins) and the non-compulsory access to the internet during the class.

FINAL REMARKS

This article aimed to propose and exemplify pedagogical uses of Voyant Tools in linguistic research introduction. Throughout the article, we presented its advantages and limitations, ten of its tools, the bases of our

20 Available at: https://products.aspose.app/pdf/conversion/srt-to-txt. Accessed: September 1, 2023.
21 Definitions available at: http://ic.votec.ileel.ufu.br/. Accessed: September 1, 2023.

pedagogical proposals, as well as three proposals for discussion rounds and two practical activities for language analysis and description.

Our proposals sought to contribute to the development of basic skills for the training of lexicon researchers, such as anticipating data, formulating and testing hypotheses based on linguistic data, methodical thinking and documentation, evaluation of the most appropriate resources and procedures, bringing together different analysis approaches, discrimination of central and peripheral data, and collaboration.

We believe that linguistic research training is relevant at all educational levels, given that it helps in the development of critical thinking and metalinguistic knowledge, both in the mother tongue and in foreign language(s). Bearing this in mind, our proposals for discussions and activities were developed considering different target audiences, such as elementary and high school students, high school students in scientific initiation, undergraduate students, undergraduate research scholars as well as graduate students. Evidently, in basic education regular classes, with pre-programmed curricula, the educational focus is not linguistic research training, but these proposals may be alternative ways of approaching aspects already programmed in the curriculum for the school year. In addition, they can also originate pedagogical and extension projects that cover the aforementioned target audiences.

Voyant Tools, despite some limitations, as all software and web environments for lexical analysis have, is a useful resource for language description research introduction, since it has a diversity of tools, an intuitive and attractive layout, as well as tools interactivity. It constitutes a more pedagogical alternative to traditionally used programs, as it allows pedagogical possibilities that other programs do not. Therefore, it is an excellent introductory resource for research in language description and Corpus Linguistics.

Following the research carried out by Fromm et al. (2020), an analytical-contrastive study of the results obtained in Voyant Tools with those obtained in other software and web environments, such as WordSmith Tools (SCOTT, 2020), AntConc (ANTHONY, 2020) and Sketch Engine (KILGARRIFF et al., 2003), is one of the possible and productive directions for future studies. Such an analysis would provide us with more information about the statistical metrics of Voyant Tools and about the reliability of the results presented by its tools.

AUTHOR CONTRIBUTION STATEMENT
Both authors have contributed equally to the writing of this paper.

CONFLICT OF INTEREST STATEMENT
The authors declare that there are no conflicts of interest.

REFERENCES

ALHUDITHI, E. (2021). Review of Voyant Tools: see through your text. *Language Learning & Technology*, v. 25, n. 3, p. 43-50. DOI: https://doi.org/10125/73446.

ANTHONY, L. (2020). *AntConc – Version 3.5.9*. Tóquio: Waseda University.

BERBER SARDINHA, T. (2004). *Lingüística de Corpus*. Barueri: Manole.

FERNÁNDEZ PÉREZ, M. (2011). El corpus *Koiné* de habla infantil: líneas maestras. In: Fernández Pérez, M. (org.). *Lingüística de Corpus y adquisición de la lengua*. Madrid: Arco Libros, p. 11-36.

FROMM, G. et al. (2020). WordSmith Tools e Sketch Engine: um estudo analítico-comparativo para pesquisas científicas com uso de corpora. *Revista de Estudos da Linguagem*, v. 28, n. 3, p. 1191-1248. DOI: http://dx.doi.org/10.17851/2237-2083.28.3.1191-1248.

FROMM, G. (2020). Por uma Terminografia Pedagógica. *Revista Estudos Linguísticos*, v. 49, n. 2, p. 761-776. DOI: https://doi.org/10.21165/el.v49i2.2637.

HENDRIGAN, H. (2019). Mixing digital humanities and applied science librarianship: using Voyant Tools to reveal word patterns in faculty research. *Issues in Science and Technology Librarianship*, n. 91, p. 1-12. DOI: https://doi.org/10.29173/istl3 .

HETENYI, G.; LENGYEL, A.; SZILASI, M. (2019). Quantitative analysis of qualitative data: using Voyant Tools to investigate the sales-marketing interface. *Journal of Industrial Engineering and Management*, v. 12, n. 3, p. 393-404. DOI: http://dx.doi.org/10.3926/jiem.2929.

JÄNICKE, S. et al. (2015). On close and distant reading in Digital Humanities: a survey and future challenges. In: Eurographics Conference On Visualization. *State-of-the-Art Reports*. Genebra: Eurographics Association, p. 83-103. DOI: http://dx.doi.org/10.2312/eurovisstar.20151113.

JOCKERS, M. L. (2011). On distant reading and macroanalysis. *Matthew L. Jockers*, 1 jul. 2011. Available at: https://www.matthewjockers.net/2011/07/01/on-distant-reading-and-macroanalysis/. Accessed: September 1, 2023.

JOCKERS, M. L. (2013). *Macroanalysis*: digital methos and literary history. Champaign: University of Illinois Press.

JOHNS, T. (1991). Should you be persuaded: two samples of data-driven learning materials. *English Language Research Journal*, v. 4, p. 1-16.

KILGARRIFF, A. et al. (2003). *Sketch Engine*. East Sussex: Lexical Computing Limited.

LISBOA, J. V. R. (2022). *Introdução ao Voyant Tools*. In: II Escola de Verão em Estudos Linguísticos. Minicurso. Uberlândia: PPGEL-UFU.

LISBOA, J. V. R.; FROMM, G. (2021). *Voyant Tools e a sala de aula*. In: XI Escola Brasileira de Linguística Computacional. Minicurso. São Paulo: FCL-Unesp/FFLCH-USP.

LORCH, M. (2016). Turning students into researchers: introduction to research methods in Applied Linguistics. *LLAS Centre for Languages, Linguistics and Area Studies*, 15 set. 2016. Available at: https://web-archive.southampton.ac.uk/www.llas.ac.uk/resources/gpg/2273.html. Accessed: September 1, 2023.

MAHONY, S.; PIERAZZO, E. (2012). Teaching skills or teaching methodology? In: Hirsch, B. D. (ed.). *Digital humanities pedagogy*: practices, principles and politics. Cambridge: Open Book Publishers, p. 215-225.

McENERY, T.; WILSON, A. (2001). *Corpus Linguistics*: an introduction. 2. ed. Edinburgh: Edinburgh University Press.

MILLER, A. (2018). Text mining digital humanities projects: assessing content analysis capabilities of Voyant Tools. *Journal of Web Librarianship*, v. 12, n. 3, p. 169-197. DOI: https://doi.org/10.1080/19322909.2018.1479673.

MORETTI, F. (2005). *Graphs, maps, trees*: abstract models for a Literary History. London: Verso.

O'KEEFFE, A.; McCARTHY, M. (ed.). (2010). *The Routledge Handbook of Corpus Linguistics*. Abingdon: Routledge, 2010. DOI: https://doi.org/10.4324/9780203856949.

OLIVEIRA, M. A.; BRITO, E. M. N.; OLIVEIRA, S. S. (2018). Diálogos sobre trabalho e saúde: análise da movimentação interativa nos blogs dos bombeiros do Rio de Janeiro, Brasil. *Ciência & Saúde Coletiva*, v. 23, n. 10, p. 3297-3307. DOI: https://doi.org/10.1590/1413-812320182310.16392018.

PERINI, M. A. (2006). *Princípios de lingüística descritiva*: introdução ao pensamento gramatical. São Paulo: Parábola.

RÖMER, U. (2008). Corpora and Language Teaching. In: Lüdeling, A.; Kytö, M. (ed.). *Corpus Linguistics*: an international handbook. Berlin: Mouton de Gruyter, p. 112-130.

SAMPSEL, L. J. (2018). Voyant Tools. *Music Reference Services Quaterly*, v. 21, n. 3, p. 153-157. DOI: https://doi.org/10.1080/10588167.2018.1496754.

SÁNCHEZ TARRAGÓ, N. (2021). Descubriendo críticas al acceso abierto mediante la visualización de textos con Voyant Tools. *Rev. Cuba. Inf. Cienc. Salud*, v. 32, n. 1, p. 1-40. Available at: http://scielo.sld.cu/pdf/ics/v32n1/2307-2113-ics-32-01-e1824.pdf. Accessed: September 1, 2023.

SCOTT, M. (2020). *WordSmith Tools – Version 8*. Stroud: Lexical Analysis Software.

SHEPHERD, T. M. G. (2012). Panorama da Linguística de Corpus. In: Shepherd, T. M. G.; Berber Sardinha, T.; Pinto, M. V. (org.). *Caminhos da Linguística de Corpus*. Campinas: Mercado de Letras, p. 15-30.

SINCLAIR, S.; ROCKWELL, G. (2012). Teaching computer-assisted text analysis: approaches to learning new methodologies. In: Hirsch, B. D. (ed.). *Digital humanities pedagogy*: practices, principles and politics. Cambridge: Open Book Publishers, p. 242-263.

SINCLAIR, S.; ROCKWELL, G. (c2023). *Voyant Tools – Version 2.6.2*.

WELSH, M. E. (2014). Review of Voyant Tools. *Collaborative Librarianship*, v. 6, n. 2, p. 96-97. Available at: https://digitalcommons.du.edu/collaborativelibrarianship/vol6/iss2/8. Accessed: September 1, 2023.

Recebido: 1/9/2023
Aceito: 20/2/2024
Publicado: 19/6/2024