



Dados Internacionais de Catalogação na Publicação (CIP)  
Sistema de Bibliotecas da UFU, MG, Brasil.

F932e Fromm, Guilherme, 1968-  
Estudos descritivos e linguística de corpus integrada à prática  
educativa (PIPE 7) / Guilherme Fromm. -- 2. ed. -- Uberlândia :  
Edufu, 2021.  
80 p. : il.

ISBN: 978-65-86084-30-6 (ebook)

Guia digital

Disponível em: <https://doi.org/10.14393/ufu-978-65-86084-30-6>

Textos em inglês e Português.

Inclui bibliografia.

1. Linguística. 2. Língua inglesa - Ensino à distância.  
3. Língua Inglesa - Guias. I. Título.

CDU 801

PRESIDENTE DA REPÚBLICA  
Jair Messias Bolsonaro

MINISTRO DA EDUCAÇÃO  
Milton Ribeiro

UNIVERSIDADE ABERTA DO BRASIL  
DIRETORIA DE EDUCAÇÃO A DISTÂNCIA/CAPES  
Carlos Cezar Modernel Lenuzza

UNIVERSIDADE FEDERAL DE UBERLÂNDIA - UFU  
REITOR  
Valder Steffen Junior

VICE-REITOR  
Carlos Henrique Martins

CENTRO DE EDUCAÇÃO A DISTÂNCIA  
DIRETOR  
Vinícius Silva Pereira

REPRESENTANTE UAB/UFU  
Maria Teresa Menezes Freitas

SUPLENTE UAB/UFU  
Aléxia Pádua Franco

INSTITUTO DE LETRAS E LINGUÍSTICA -ILEEL -  
UFU  
DIRETOR  
Ariel Novodvorski

CURSO DE LETRAS - LICENCIATURA EM  
INGLÊS E LITERATURAS DE LÍNGUA INGLESA  
COORDENADOR  
Rafael Matielo

ILEEL – UFU CONSELHO EDITORIAL  
Aléxia Pádua Franco - UFU  
Bruno Franceschini - UFG

Diva Souza Silva - UFU  
Maria Teresa Menezes Freitas - UFU  
Simone Tiemi Hashiguti – UFU  
Stella Esther Ortweiler Tagnin - USP  
Viviane Cabral Benzegen – UFV

**EQUIPE DO CENTRO DE EDUCAÇÃO A  
DISTÂNCIA DA UFU - CEaD/UFU**

ASSESSORA DA DIRETORIA  
Sarah Mendonça de Araújo

EQUIPE MULTIDISCIPLINAR  
Alberto Dumont Alves Oliveira  
Darcus Ferreira Lisboa Oliveira  
Dirceu Nogueira de Sales Duarte Júnior  
Gustavo Bruno do Vale  
Otaviano Ferreira Guimarães

## SUMÁRIO

SUMÁRIO .....	4
FIGURAS.....	5
TABELAS .....	7
INFORMAÇÕES .....	8
SOBRE O AUTOR.....	9
INTRODUÇÃO.....	10
AGENDA .....	11
<b>MÓDULO 1 .....</b>	<b>12</b>
<i>Basic principles of Corpus Linguistics</i> .....	13
ACTIVITY 1 .....	13
ACTIVITY 2.....	13
<i>Corpus Linguistics?</i> .....	13
ACTIVITY 3 - SEARCH.....	14
ACTIVITY 4 - DISCUSSION .....	14
<i>Corpus Linguistics in History</i> .....	16
<i>Corpora Typology and Planning</i> .....	18
<i>Key-Concepts</i> .....	20
<b>MÓDULO 2 .....</b>	<b>29</b>
<i>Corpus Linguistics: Teaching and Learning</i> .....	30
ACTIVITY 5.....	30
<i>At school, at home, self-learning</i> .....	30
ACTIVITY 6 - SEARCH.....	34
ACTIVITY 7 - DISCUSSION .....	34
ACTIVITY 8 - PIPE .....	38
<b>MÓDULO 3 .....</b>	<b>39</b>
<i>Lexical Analysis Tools</i> .....	40
ACTIVITY 9.....	40
<i>Compiling a corpus</i> .....	42
ACTIVITY 10 - CORPUS COMPILATION .....	47
<i>AntConc</i> .....	47
<b>MÓDULO 4 .....</b>	<b>67</b>
<i>Project Development</i> .....	68
ACTIVITY 11 .....	68
ACTIVITY 12.....	68
<i>Describing language nowadays</i> .....	68
<i>Choosing an area to be described: Terminography</i> .....	69
<i>Web environment for terminological management: VoTec</i> .....	71
References .....	79
Suggested books.....	80



## FIGURAS

	Availability	Page
<b>Figure 1</b>	Source: YAMAMOTO, Márcio Issamu. VoBLing: vocabulário bilíngue de linguística, português-inglês, direcionado por corpus. 2020. 214 f. Tese (Doutorado em Estudos Linguísticos) - Universidade Federal de Uberlândia, Uberlândia, 2020. Disponível em: <a href="http://doi.org/10.14393/ufu.te.2020.682">http://doi.org/10.14393/ufu.te.2020.682</a> .	<b>14</b>
<b>Figure 2</b>	Source: from the author.	<b>15</b>
<b>Figure 3</b>	Source: from the author.	<b>16</b>
<b>Figure 4</b>	Source: from the author.	<b>16</b>
<b>Figure 5</b>	Source: from the author.	<b>17</b>
<b>Figure 6</b>	Source: from the author. Wordlist of the Linguistics Corpus.	<b>21</b>
<b>Figure 7</b>	Statistics. Source: from the author.	<b>22</b>
<b>Figure 8</b>	Keywords for the Sociolinguistics subcorpus. Source: from the author.	<b>23</b>
<b>Figure 9</b>	Concordance lines for the word VALENCE in the Linguistics Corpus. Source: from the author.	<b>24</b>
<b>Figure 10</b>	Text tab for the word VALENCE; line 16 (valence alternation) chosen. Source: from the author.	<b>25</b>
<b>Figure 11</b>	Collocates for the word VALENCE. Source: from the author.	<b>26</b>
<b>Figure 12</b>	Concordance lines with the words VALENCE and SUFFIXES. Source: from the author.	<b>27</b>
<b>Figure 13</b>	Cluster tab for the word VALENCE. Source: from the author.	<b>28</b>
<b>Figure 14</b>	Concordance lines for the cluster verb root valence. Source: from the author.	<b>28</b>
<b>Figure 15</b>	Registration corner. Source: COCA. Available at: <a href="https://www.english-corpora.org/coca/">https://www.english-corpora.org/coca/</a>	<b>30</b>
<b>Figure 16</b>	COCA's upper menu. Available at: <a href="https://www.english-corpora.org/coca/">https://www.english-corpora.org/coca/</a>	<b>31</b>
<b>Figure 17</b>	COCA's search display. Source: COCA. Available at: <a href="https://www.english-corpora.org/coca/">https://www.english-corpora.org/coca/</a>	<b>32</b>
<b>Figure 18</b>	COCA's query result. Source: COCA. Available at: <a href="https://www.english-corpora.org/coca/">https://www.english-corpora.org/coca/</a>	<b>32</b>
<b>Figure 19</b>	COCA's concordance lines for the query at the Internet. Source: COCA. Available at: <a href="https://www.english-corpora.org/coca/">https://www.english-corpora.org/coca/</a>	<b>33</b>
<b>Figure 20</b>	Expanded concordance line for the first example. Source: COCA. Available at: <a href="https://www.english-corpora.org/coca/">https://www.english-corpora.org/coca/</a>	<b>33</b>
<b>Figure 21</b>	Word and Phrase main page. COCA platform. Source: COCA. Available at: <a href="https://www.english-corpora.org/coca/">https://www.english-corpora.org/coca/</a>	<b>35</b>
<b>Figure 22</b>	Frequency query for Linguistics. Source: COCA. Available: <a href="https://www.english-corpora.org/coca/">https://www.english-corpora.org/coca/</a>	<b>36</b>
<b>Figure 23</b>	Word LINGUISTICS description. Available at: <a href="https://www.english-corpora.org/coca/">https://www.english-corpora.org/coca/</a>	<b>37</b>
<b>Figure 24</b>	Menu to get to new screens. Source: COCA. Available: <a href="https://www.english-corpora.org/coca/">https://www.english-corpora.org/coca/</a>	<b>37</b>
<b>Figure 25</b>	Text insertion. Available at: <a href="https://www.english-corpora.org/coca/">https://www.english-corpora.org/coca/</a>	<b>37</b>
<b>Figure 26</b>	Text analysis. Available at: <a href="https://www.english-corpora.org/coca/">https://www.english-corpora.org/coca/</a>	<b>38</b>
<b>Figure 27</b>	Corpora do Português, COCA platform. Source: COCA. Available: <a href="http://www.corpusdoportugues.org/x.asp">http://www.corpusdoportugues.org/x.asp</a>	<b>38</b>
<b>Figure 28</b>	WordSmith Tools' main menu. Source: from the author.	<b>40</b>
<b>Figure 29</b>	WordSmith Tools utilities. Source: from the author.	<b>41</b>
<b>Figure 30</b>	AntConc's main menu. Source: from the author.	<b>42</b>
<b>Figure 31</b>	Google: query for the word <i>Prosody</i> . Source: from the author.	<b>43</b>
<b>Figure 32</b>	File <i>Prosody: Rhythms and Melodies of Speech</i> first page. Source: <a href="https://arxiv.org/ftp/arxiv/papers/1704/1704.02565.pdf">https://arxiv.org/ftp/arxiv/papers/1704/1704.02565.pdf</a> .	<b>44</b>
<b>Figure 33</b>	Path to open the Notepad program. Source: from the author.	<b>44</b>
<b>Figure 34</b>	File selected. Source: from the author.	<b>45</b>
<b>Figure 35</b>	File copied to the Notepad program.	<b>45</b>
<b>Figure 36</b>	Header insertion. Source: from the author.	<b>46</b>
<b>Figure 37</b>	Saving the file as File 1. Source: from the author	<b>46</b>
<b>Figure 38</b>	AntConc. Opening a directory in the WordList tool. Source: from the author.	<b>47</b>
<b>Figure 39</b>	AntConc. Wordlist and files. Source: from the author.	<b>48</b>
<b>Figure 40</b>	AntConc. Corpus of Prosody, processed by the Word List tool. Source: from the author.	<b>49</b>

<b>Figure 41</b>	AntConc. Tool Preferences menu. Source: from the author.	<b>50</b>
<b>Figure 42</b>	Word List with Stoplist, Prosody. Source: from the author.	<b>51</b>
<b>Figure 43</b>	Loading the Lemma List. Source: from the author.	<b>52</b>
<b>Figure 44</b>	Word List with Stoplist and Lemma list. Source: from the author.	<b>53</b>
<b>Figure 45</b>	Saving file in AntConc. Source: from the author.	<b>54</b>
<b>Figure 46</b>	File saved from AntConc's Word List. Source: from the author.	<b>54</b>
<b>Figure 47</b>	Concordance lines for the word PRAGMATIC. Source: from the author.	<b>55</b>
<b>Figure 48</b>	File View for the first entry of the word PRAGMATIC. Source: from the author.	<b>55</b>
<b>Figure 49</b>	Concordance Plot for the word PRAGMATIC. Source: from the author.	<b>56</b>
<b>Figure 50</b>	Concordance lines for the word PRAGMATICS. Source: from the author.	<b>57</b>
<b>Figure 51</b>	Kwic sort configurations for the word PRAGMATICS. Source: from the author.	<b>57</b>
<b>Figure 52</b>	Searching for the specific word SPEAKER in the WordList. Source: from the author.	<b>58</b>
<b>Figure 53</b>	Collocates for the word SPEAKER. Source: from the author.	<b>59</b>
<b>Figure 54</b>	Collocates for SPEAKER + HEARER. Source: from the author.	<b>60</b>
<b>Figure 55</b>	Clusters for the word SPEAKER. Source: from the author.	<b>61</b>
<b>Figure 56</b>	Clusters for the word SPEAKER. Source: from the author.	<b>62</b>
<b>Figure 57</b>	SPEAKER'S UTTERANCE clusters. Source: from the author.	<b>63</b>
<b>Figure 58</b>	AntConc's Tool Preferences and Keyword List. Source: from the author.	<b>64</b>
<b>Figure 59</b>	Keyword List from the PRAGMATICS corpus. Source: from the author.	<b>65</b>
<b>Figure 60</b>	Concordance lines for the word DISCOURSE. Source: from the author.	<b>66</b>
<b>Figure 61</b>	Concordances lines for the combination PRAGMATICS + IS. Source: from the author.	<b>70</b>
<b>Figure 62</b>	Concordance for the combination PRAGMATICS + :. Source: from the author.	<b>70</b>
<b>Figure 63</b>	Main consulting page of VoTec (in English), Linguistics area. Source: from the author.	<b>71</b>
<b>Figure 64</b>	VoTec researcher's first page. Source: from the author.	<b>72</b>
<b>Figure 65</b>	Creating the term PROSODY. Source: from the author.	<b>72</b>
<b>Figure 66</b>	Inserting a context for PROSODY. Source: from the author.	<b>73</b>
<b>Figure 67</b>	Inserted contexts for PROSODY. Source: from the author.	<b>74</b>
<b>Figure 68</b>	Second page for the term PROSODY. Source: from the author.	<b>74</b>
<b>Figure 69</b>	Data tab. Source: from the author.	<b>75</b>
<b>Figure 70</b>	Distinctive Traces tab. Source: from the author.	<b>75</b>
<b>Figure 71</b>	Semantics tab. Source: from the author.	<b>76</b>
<b>Figure 72</b>	Equivalent Term tab. Source: from the author.	<b>76</b>
<b>Figure 73</b>	Cross-reference terms tab. Source: from the author.	<b>77</b>
<b>Figure 74</b>	Encyclopedic Information tab. Source: from the author.	<b>77</b>
<b>Figure 75</b>	Final Concept and Definition tabs. Source: from the author.	<b>77</b>
<b>Figure 76</b>	Term PROSODY in VoTec's page (just English). Source: from the author.	<b>78</b>
<b>Figure 77</b>	Term GRAMMAR in the VoTec. Source: from the author.	<b>78</b>

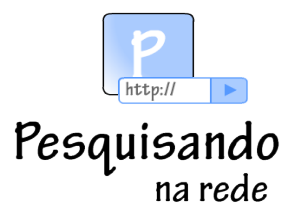
## TABELAS

<b>Table 1. Collaborative Corpus of Linguistics typology .....</b>	<b>18</b>
<b>Table 2. Features related to COCA. ....</b>	<b>31</b>
<b>Table 3. Query possibilities using the COCA interface. ....</b>	<b>35</b>
<b>Table 4. Corpus compilation .....</b>	<b>42</b>

## INFORMAÇÕES

Prezado(a) aluno(a),

Ao longo deste guia impresso você encontrará alguns “ícones” que lhe ajudará a identificar as atividades.



Fique atento ao significado de cada um deles, isso facilitará a sua leitura e seus estudos.

Destacamos alguns termos no texto do Guia cujos sentidos serão importantes para sua compreensão. Para permitir sua iniciativa e pesquisa não criamos um glossário, mas se houver dificuldade interaja no *Fórum de Dúvidas*.

## **SOBRE O AUTOR**

Guilherme Fromm é professor de Língua Inglesa do Instituto de Letras e Linguística da Universidade Federal de Uberlândia. cursou História e Letras (Alemão/Português) na graduação. Especializou-se em Tradução (Inglês/Português), cursou o mestrado na área de Linguística, o doutorado na área de Letras - Estudos Linguísticos e Literários em Inglês e o pós-doutorado na área do Léxico. Toda sua formação acadêmica se deu na Universidade de São Paulo, com exceção do pós-doutorado, realizado na UFSCar.

Atuou como professor de inglês em cursos livres por quinze anos e há dezenove é professor universitário, nas áreas de Linguística e Língua Inglesa. Diretor das revistas Domínios de Lingu@gem (<http://www.seer.ufu.br/index.php/dominiosdelinguagem>), GTLex (<http://www.seer.ufu.br/index.php/GTLex/index>) e da série de livros e-Classe (pela EDUFU) desde o seu início. É autor de artigos e capítulos de livros e organizador de livros. Tem experiência na área de Linguística, atuando nos seguintes temas: Ciências do Léxico (Lexicologia, Lexicografia, Terminologia, Terminografia), Linguística de Corpus, Ensino de Língua Inglesa e Tradução.

## INTRODUÇÃO

Olá! Seja bem-vindo.

Com este guia, trabalharemos uma disciplina diferenciada dentro do curso PARFOR/ Inglês. O objetivo geral é aprender a descrever língua, ao invés de só ensiná-la. Para tanto, adotaremos a abordagem e a metodologia da Linguística de *Corpus*.

Embora a ênfase do nosso trabalho, aqui, seja a descrição, também mostraremos como ferramentas da Linguística de *Corpus* podem ser trabalhadas na sala de aula para o enriquecimento da experiência de aprendizagem do aluno.

Como trabalho prático para o PIPE, compilaremos um *corpus* e prepararemos um plano de aula detalhado usando as várias possibilidades de atividades aqui expostas.



## AGENDA

WEEK	MODULES	STUDY DEVELOPMENT	EVALUATION
Weeks 1 and 2	<b>Module 1</b> Basic principles of Corpus Linguistics	Activity 1 – chat with the tutor. Activity 2 – video class. Activity 3 – Corpus Linguistics search. Activity 4 – Forum discussion	Activity 4 Score: 10 points.
Weeks 3 and 4	Module 2 Corpus Linguistics: Teaching and Learning	Activity 5 – video class. Activity 6 – search using COCA. Activity 7 – Forum discussion Activity 8 – PIPE. Search and describe a site that uses corpora and tools to analyze language.	Activity 7 Score: 5 points. Activity 8 Score: 10 points.
Weeks 5 and 6	<b>Module 3</b> Lexical Analysis Tools	Activity 9 – video class. Activity 10 – Corpus Compilation. You're going to compile a study corpus.	Activity 10 Score: 10 points.
Weeks 7 and 8	<b>Module 4</b> Project Development	Activity 11 – video class. Activity 12 – PIPE. You're going to prepare a class plan project, using the Corpus Linguistics approach.	Activity 12 Score: 15 points.

# MÓDULO 1

## Basic principles of Corpus Linguistics

### **Basic Contents**

- What is Corpus Linguistics?
- What to consider?
- Corpus Linguistics in History.
- Corpora typology and planning.
- Key-concepts: frequency, keywords, concordances, collocations, clusters/n-grams.

### **Objectives**

- Delimitate the activity field.
- Know that the practice is old, but the way we work today is new.
- Decide the ways you implement a research.
- Know the importance in corpus design and how to identify it.
- Understand the basic concepts related to Corpus Linguistics.

## BASIC PRINCIPLES OF CORPUS LINGUISTICS



### ACTIVITY 1

Chat with your tutor. Take a look at the site (AVA) the date and schedule of this chat.



### ACTIVITY 2

Video class, module 1. Watch the professor's hints about the subjects that are going to be worked in this module.

## Corpus Linguistics?

First, we need to delimitate the field we are studying. What is Corpus Linguistics (CL from now on)? Let us see a definition provided by the CASS Centre:

Corpus linguistics, broadly, is a collection of methods for studying language. It begins with collecting a large set of language data – a corpus -, which is made usable by computers. Corpora (the plural of corpus) are usually so large that it would be impossible to analyse them by hand, so software packages (often called concordancers) are used in order to study them. It is also important that a corpus is built using data well matched to a research question it is built to investigate. To investigate language use in an academic context, for example, it would be appropriate for one to collect data from academic contexts such as academic journals or lectures. Collecting data from the sports pages of a tabloid newspaper would make much less sense.



An important idea we must consider is that CL, nowadays, means the usage of a computer. This is a kind of **empiric** (based on facts) work.



### ACTIVITY 3 - SEARCH

What about you finding other concepts of CL available on the Internet? Do a search using Google (or other searcher) about the term. What have you found interesting about the theme?



### ACTIVITY 4 - DISCUSSION

Discuss with the other students, using the AVA forum, which concepts about CL you considered interesting. Try to identify possible approaches for this subject. The participation in the discussion will be evaluated.

### What to consider?

When working with research, you must have in mind that there's a division among theory

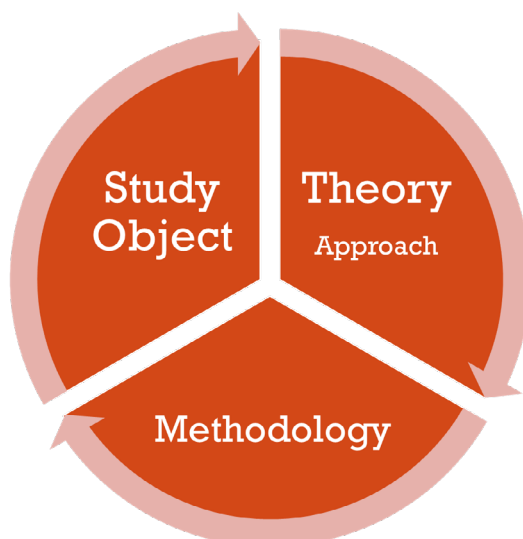


Figure 1. The process of analyzing science.

Just in our area, we have a lot of study fields:



Figure 2. Linguistics Domain Tree.

Theories:



Figure 3. Some linguistic Theories.

And methodologies (this is a tiny example, there are more methodologies):

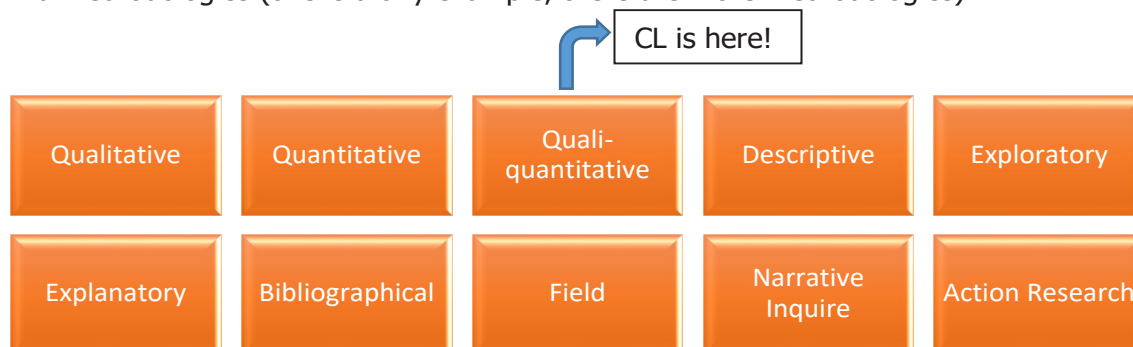


Figure 4. Some Methodologies.

## Corpus Linguistics in History

There's nothing new about collecting corpora for studies. Researchers have been doing this for ages. According to Berber Sardinha:

Havia corpora antes do computador, já que o sentido original da palavra corpus é corpo, conjunto de documentos (conforme o dicionário Aurélio). Na Grécia Antiga, Alexandre, o Grande definiu o Corpus Helenístico. Na Antiguidade e na Idade Média, produziam-se corpora de citações da Bíblia.

Many corpora in English language have been developed since the advent of the computer age. From the early modern corpora (like Brown Corpus), through those who were a mark (like BNC) up to the most contemporary ones (like COCA), the size differ a lot. The Brown Corpus had 1 million words, the BNC had 100 million, and the COCA has 450

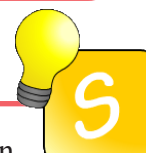


million (up to the moment). There are corpora that oversize one billion words, like the GloWbe (Global Web-Based English), with 1.9 billion. Take a look here (<http://corpus.byu.edu/>) and compare sizes.

What is new about the way we're studying today is the approach we have: we develop a research project, from the scratches to the results, using a computer. That's the reason we can say that **Corpus Linguistics is as an approach as well as a methodology** to be used in linguistic studies.



We can't mention the history of Corpus Linguistics without Chomsky. When developed the Generative Grammar first concepts (with the ideas of competence and performance), he criticized a lot the researchers who were using corpora for analysis.



The Generative Grammar and other Chomskyan concepts are particularly important in Linguistics history. Let's learn about these concepts? Search the net and write some of them down.



Not all kinds of theories and methodologies are compatible. When we use the Generativist theory, for example, we can't use the CL methodology, since they're not compatible. On the other hand, although CL seeks the idea how you perform language (we study the way languages are really used), the computational data set of CL lexical analysis programs are based on the Generative Grammar (the theoretical competence of a determined language). A paradox!



**Generativism**  
**competence**



**Be careful with the choices!**



**Corpus Linguistics**  
**performance**

Computational Paradox

Figure 5. Comparing Generativism and Corpus Linguistics.

## Corpora Typology and Planning

Corpus compilation is something that must be carried out very carefully. Imagine you've compiled a study corpus for a project and you discover, later, that the data you have collected is not big enough or not balanced enough throughout the areas you wish to study, or the sources you used are not trustworthy. The whole set of results that come from the corpus, in this case, can be invalidated.

See the table below. It's a summary of the way a corpus was organized.

Table 1. Collaborative Corpus of Linguistics typology. Source:  
Fromm, Yamamoto (2013)

Língua	Bílingue (Inglês e Português)
Modo	Escrito (textos acadêmicos: artigos científicos, dissertações e teses)
Data de Publicação	Sincrônico (levantamento realizado entre 2010 e 2014)
Seleção	Amostragem, Estático
Conteúdo	Especializado (Linguística)
Autoria	Falantes nativos/não nativos (inglês e português), individual/coletivo
Disposição Interna	Comparável
Uso na Pesquisa	Estudo (Análise terminológica/terminográfica)
Tamanho	Grande (mais de 10 milhões de palavras)
Nível de Codificação	Sem cabeçalhos, sem etiquetas

Let's take a look at each of these possibilities in corpus designing?

### *Language*

Depending on your project, you can analyze one, two or multiple languages (for a contrastive work, for example). The example from table 1 works with two languages (English and Portuguese).

### *Sources*

Basically, we have two sources corpora are compiled from: written sources or oral sources. Written corpora are very easily found on the Internet, can be scanned from books, etc. Oral corpora are more difficult to be worked with: the programs used for Corpus Linguistics don't analyze the sounds of recordings; we must transcribe the sounds into a written file first, and then analyze them.

Nevertheless, other combinations can be arranged. Beilke<sup>1</sup>, for examples, intends to use the texts from gravestones to analyze the Pommern dialect used in Brazil by some

<sup>1</sup> BEILKE, Neubiana Silva Veloso. Pommersche Korpora: uma proposta metodológica para compilação de corpora dialetais. 2016. 285 f. Dissertação (Mestrado em Estudos Linguísticos) - Universidade Federal de Uberlândia, Uberlândia, 2016. Disponível em: <http://doi.org/10.14393/ufu.di.2016.426>

communities that immigrated from Germany and Poland. It's an unusual source of a written corpus. In addition, she also intends to analyze traditional written (from clerical sources) and oral (from interviews) corpora as sources.

### *Time*

You can try to compile a diachronic corpus or a synchronic one. A diachronic corpus is much more difficult to be prepared, because it generally involves scanning work.

### *Selection*

Great corpora of a language, like the COCA, are called general or reference corpora. The corpus you prepare for your research, much smaller, is called sample or study corpus. If you prepare your corpus and finish it, it's called a static corpus. If the corpus is continuously changed, we name it dynamic corpus. All these previous possibilities imply the idea of **balance**: the amount of texts or numbers of words must be very well distributed among the corpora genres or subsections.

### *Content*

A general corpus must contain, theoretically, all the traces of a language; it's very difficult to design one, since a deep previous study must be done before you start collecting texts. A study corpus contains texts that are related to the study you're developing in your research.

### *Authorship*

The texts you collect for your corpus may be written/spoken by native or non-native speakers of a language. These texts can also be produced individually or in a group (more than one author/speaker).

### *Internal distribution*

For bilingual or multilingual works, the corpora you prepare are parallel or comparable. For example, if you compile texts in English and their translations into Portuguese, these corpora are *parallel*. On the other hand, you can choose an area for your study and take texts you find about that area in each language; they're not the same texts, but belong to the same area – in this case, they are *comparable*.

### *Size*

Berber Sardinha proposes a table to classify the size of a corpus. It's a little bit tricky, however, since nowadays anyone can download hundreds of texts in minutes using the Internet. Let's say that, obeying the other parameters we comment here, the bigger the

size of the corpus the best it is to represent the language (specialized or not).

### *Codification level*

For more specific kinds of studies, the corpora may contain tags (like the morphosyntactic ones, in which each word of a corpus is classified according to its word classes or syntactic

**Lingüística de Corpus**, from Berber Sardinha, is a very important book in Corpus Linguistics area in Brazil. It contains all the basic research principles for the area. Read it!



### **Key-Concepts**

There are some basic concepts involving the Corpus Linguistics approach that we must always have in mind. We're going to present some of them using screens of the WordSmith Tools (version 6<sup>2</sup>) program. For our work with lexical analysis software in module 3, however, we're going to work with AntConc, since it's a free software.

### *Frequency*

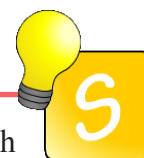
Frequency is the basis of the Corpus Linguistics. Remember we're working in an empiricist way, so the facts we analyze are based on numbers and statistics. The first kind of a work you do with your corpus is to process it through a wordlist tool. See picture 1 for an example of a wordlist.

<sup>2</sup> The lexical analysis suite is already in version 8.

N	Word	Freq.	%	Texts	%	L	§
1	THE	1.385.103	5.17	1.557	99.81		
2	OF	888.380	3.31	1.558	99.87		
3	AND	665.623	2.48	1.557	99.81		
4	A	629.708	2.35	1.555	99.68		
5	IN	601.361	2.24	1.556	99.74		
6	TO	542.813	2.02	1.557	99.81		
7	IS	312.029	1.16	1.549	99.29		
8	L	274.177	1.02	1.100	70.51		
9	THAT	266.795	1.00	1.536	98.46		
10	I	236.447	0.88	1.404	90.00		
11	E	226.330	0.84	1.388	88.97		
12	FOR	208.442	0.78	1.544	98.97		
13	AS	206.614	0.77	1.543	98.91		
14	S	177.989	0.66	1.404	90.00		
15	LANGUAGE	153.163	0.57	1.477	94.68		
16	ARE	149.951	0.56	1.534	98.33		
17	THIS	143.674	0.54	1.533	98.27		
18	BE	142.669	0.53	1.529	98.01		
19	N	138.344	0.52	1.072	68.72		
20	T	137.151	0.51	1.193	76.47		
21	WITH	134.872	0.50	1.548	99.23		
22	ON	132.258	0.49	1.537	98.53		
23	R	129.867	0.48	1.155	74.04		
24	IT	124.325	0.46	1.525	97.76		
25	OR	115.875	0.43	1.528	97.95		
26	BY	114.474	0.43	1.540	98.72		
27	O	112.602	0.42	766	49.10		
28	NOT	103.454	0.39	1.506	96.54		
29	C	93.841	0.35	1.223	78.40		
30	AN	93.170	0.35	1.522	97.56		
31	FROM	90.602	0.34	1.531	98.14		
32	WHICH	89.464	0.33	1.504	96.41		
33	D	81.837	0.31	1.261	80.83		
34	P	81.438	0.30	1.189	76.22		
35	M	79.686	0.30	1.237	79.29		
36	HAVE	75.897	0.28	1.514	97.05		
37	G	72.238	0.27	1.288	82.56		
38	THEIR	68.306	0.25	1.478	94.74		
39	AT	66.332	0.25	1.510	96.79		
40	WAS	66.311	0.25	1.398	89.62		
41	THEY	66.231	0.25	1.480	94.87		
42	CAN	64.349	0.24	1.495	95.83		

Figure 6. Wordlist of the Linguistics Corpus.

This first screen of the Wordlist tools emphasizes the frequency (tab 1). As you can see, the article THE is the first one. It appears 1.385.103 times in the corpus texts; it represents 5.17% of all the words in the texts; it appears in 1.557 texts, which represent 99.81% of the texts of the corpus.



As you can see, the grammar words (prepositions, articles, pronouns, etc.) are much more frequent than content words (nouns, verbs, adverbs, adjectives). Usually, the article THE is the most frequent word in English.

If we want more information about the corpus, we can go to the statistics tab (1). Take a look at figure 7:

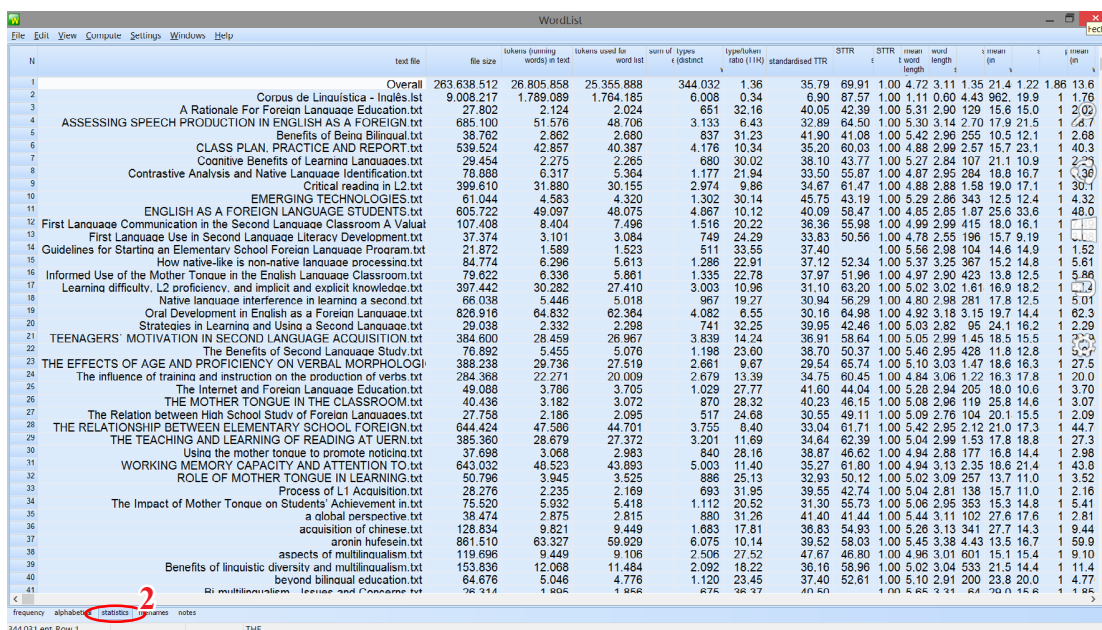
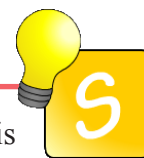


Figure 7. Statistics

In the statistics tab (2) we can find: the name of the text that belong to the corpus (text file); the size of the corpus in bytes: approx. 263.63 Mb; the size of the corpus in total words (or tokens): 26.805.858 words; the size of the corpus used for the wordlist: 25.355.888 words<sup>3</sup>; the amount of distinct words (types<sup>4</sup>): 344.032 different words; the token/type ratio<sup>5</sup>; etc.



When we're talking about Corpus Linguistics, we're talking about computers. This also means we're talking about statistics, programming, Computational Linguistics, mathematics, etc. If you want to get deeper in this area, studying these topics is surely a good idea.

<sup>3</sup> By default, the programs don't count numbers (as words) for analysis.

<sup>4</sup> As you can see in fig. 1, the article THE appears 1.385.103 times (tokens) in the corpus, but it's just one word (type).

<sup>5</sup> You divide one by another. The idea is the following: the bigger the number of the ratio, the lexically denser the corpus is. Theoretically, a denser text is more difficult to be read.



## Keywords

With the keywords tool, we can find out the words that are key to one area of study. In figure 8, we have the keyword list for the Sociolinguistics subcorpus of the Linguistics corpus.

N	Key word	Freq.	%	Texts	RC. Freq.	RC. %	Keyness	P	L	S
1	LANGUAGE	5.249	1.03	??	22.329	0.02	30.930.87	0.00		
2	HITTITE	1.908	0.37	??	23		20.672.42	0.00		
3	LUVIAN	1.681	0.33	??	1		18.415.04	0.00		
4	LINGUISTIC	1.744	0.34	??	2.772		13.121.51	0.00		
5	SOCIOLOGUÍSTICS	1.015	0.20	??	67		10.626.01	0.00		
6	DIALECT	1.168	0.23	??	709		10.322.98	0.00		
7	SOCIOLOGUÍSTIC	639	0.12	??	129		6.311.03	0.00		
8	SPEAKER	1.074	0.21	??	8.830		5.049.57	0.00		
9	CTH	363	0.07	??	2		3.954.50	0.00		
10	ANATOLIAN	360	0.07	??	51		3.638.59	0.00		
11	SPEECH	859	0.17	??	9.811		3.523.92	0.00		
12	STUDY	1.239	0.24	??	31.553	0.03	3.298.99	0.00		
13	ED	591	0.12	??	3.192		3.226.80	0.00		
14	VARIATION	621	0.12	??	3.895		3.223.77	0.00		
15	KUB	278	0.05	??	0		3.047.46	0.00		
16	MELCHERT	277	0.05	??	0		3.036.50	0.00		
17	SPEAK	784	0.15	??	10.487		2.990.70	0.00		
18	ENGLISH	1.090	0.21	??	27.147	0.02	2.944.53	0.00		
19	TEXT	711	0.14	??	8.527		2.853.34	0.00		
20	CODE	643	0.13	??	6.522		2.776.67	0.00		
21	ZA	302	0.06	??	133		2.776.06	0.00		
22	SOCIAL	1.302	0.25	??	45.508	0.04	2.758.84	0.00		
23	LDS	269	0.05	??	68		2.610.44	0.00		
24	ARZAWA	236	0.05	??	0		2.587.04	0.00		
25	KIZZUWATNA	235	0.05	??	0		2.576.07	0.00		
26	WORD	891	0.17	??	23.889	0.02	2.291.91	0.00		
27	FORM	1.085	0.21	??	38.429	0.03	2.274.57	0.00		
28	WA	315	0.06	??	623		2.260.97	0.00		
29	HATTUSA	206	0.04	??	1		2.245.51	0.00		
30	PUNJABI	234	0.05	??	72		2.231.81	0.00		
31	KBO	203	0.04	??	0		2.225.28	0.00		
32	STANDARD	756	0.15	??	16.510	0.01	2.217.08	0.00		
33	VARIETY	633	0.12	??	10.415		2.177.11	0.00		
34	SWITCH	459	0.09	??	3.803		2.151.09	0.00		
35	LABOV	239	0.05	??	141		2.119.87	0.00		
36	FISHMAN	207	0.04	??	34		2.073.27	0.00		
37	WRITE	656	0.13	??	12.926	0.01	2.043.81	0.00		
38	ANATOLIA	218	0.04	??	109		1.974.34	0.00		
39	FEATURE	530	0.10	??	7.587		1.956.27	0.00		
40	CF	291	0.06	??	827		1.914.85	0.00		
41	GENITIVE	190	0.04	??	30		1.907.76	0.00		
42	SCRIBE	214	0.04	??	155		1.845.09	0.00		
43	BAKISTAN	0.10	0.02	??	0.007		1.700.10	0.00		

Ms plot links clusters filenames source text notes

87 entries Row 1 LANGUAGE

Figure 8. Keywords for the Sociolinguistics subcorpus.

In this case, the list starts with the most key word in the corpus, according to its keyness<sup>6</sup> (3). The example shows that the word LANGUAGE is the most key in the corpus; you can see that its frequency is 5.249 entries and its keyness is 30.930,87. Keyness is not related to frequency. If you compare the words SOCIOLOGICAL and DIALECT, you can notice that DIALECT is more frequent than SOCIOLOGICAL, but the keyness values are inverted.

Not all researchers use the Keyword tool for their researches. Nevertheless, it's a very useful tool for terminological works, the kind of work we're going to present in module 4.

### Concordance

You can choose a word, from the Wordlist or Keywords screens, and ask for its concordance lines. The result, in a KWIC (keyword in context) screen, shows all the times the chosen word appears in the corpus, in example lines taken directly from the texts, and with the chosen word centralized in a different color. Let's observe the concordances for the word VALENCE (figure 9):

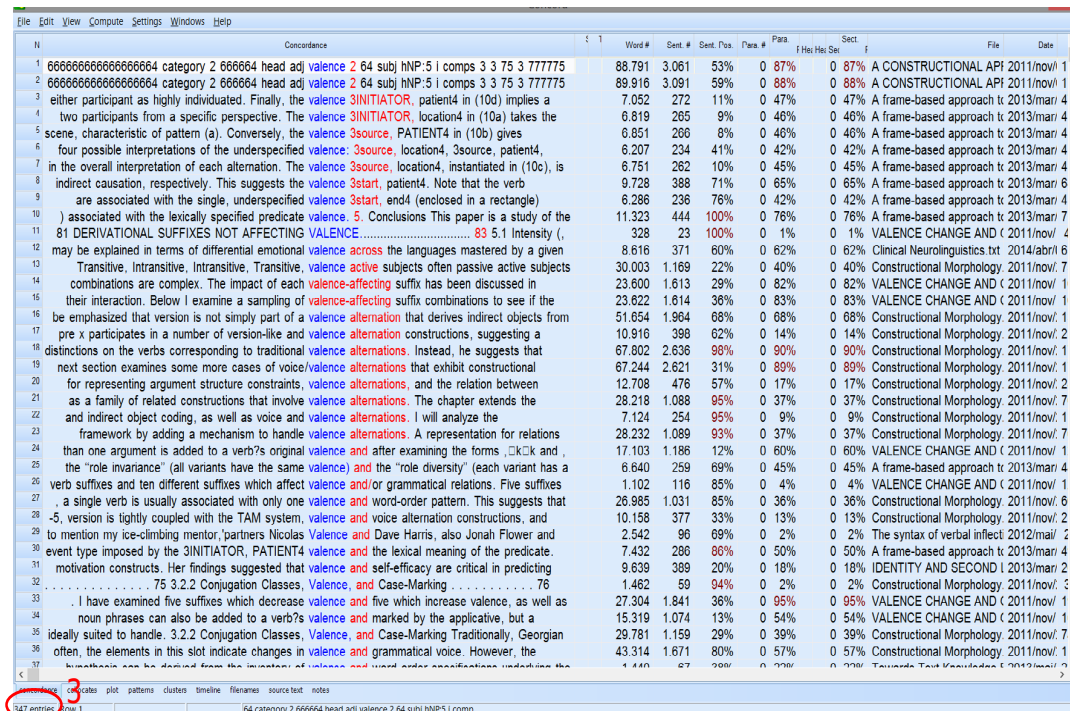


Figure 9. Concordance lines for the word VALENCE in the Linguistics Corpus.

As you can see, the chosen word (VALENCE) is centralized and its color is different (blue). The first word to the right of VALENCE (in red) is classified alphabetically. In this example, we can find 347 lines of concordances (4). In WordSmith Tools, if you click twice in the line you want to analyze, the text is going to be open in other tab, with that word highlighted. Let's analyze the line 16 of picture 4 by double clicking it (figure 10):

<sup>6</sup> Keyness is how key the word is. To get to this number, the program compares your study corpus with a reference corpus (which you must get previously) and statistically calculates the probabilities of usage in each one.

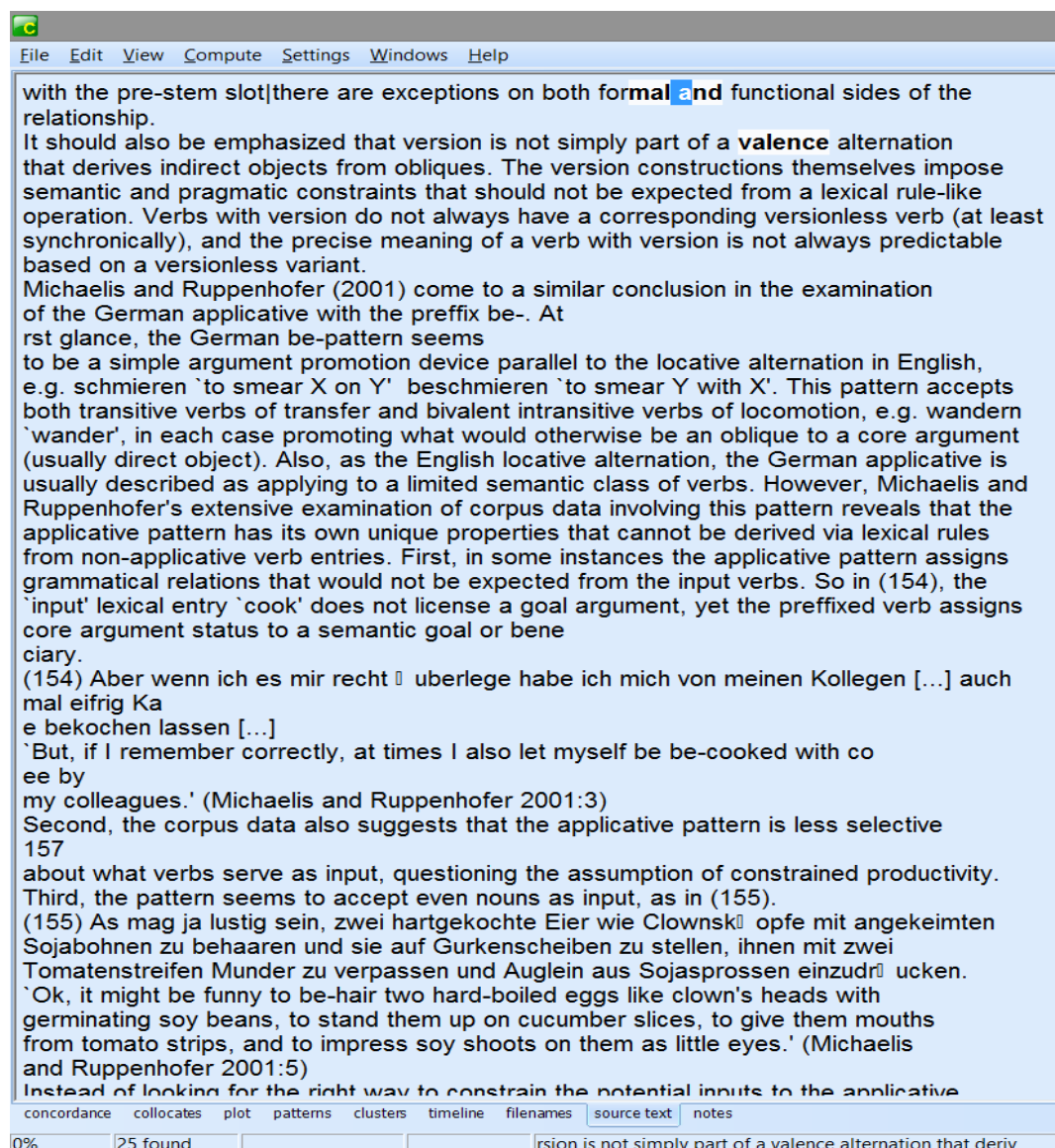


Figure 10. Text tab for the word VALENCE; line 16 (valence alternation) chosen.

As you can see, we have the chosen example (valence alternation) highlighted in the second paragraph of the text. The text can be copied using the control C command.

Generally, the concordance lines are the starting point of your research. It's from here that you're going to identify linguistic phenomena that are pertinent for your studies.

### *Collocates*

According to McGlashan (2013), collocates are [...] "A co-occurrence relationship between words or phrases. Words are said to collocate with one another if one is more likely to occur in the presence of the other than elsewhere." If we analyze the collocates tab in the VALENCE concordance screen, we get the following results (Figure 11):

N	Word	With	Relation	Texts	Total	Total Left	Total Right	L5	L4	L3	L2	L1	Centre	R1	R2	R3	R4	R5
1	VALENCE	valence	0,000	38	353	4	4	1	2		1		345		1		2	1
2	THE	valence	0,000	24	204	137	67	15	25	19	15	63		3	20	16	14	14
3	OF	valence	0,000	25	105	52	53	5	5	7	16	19		30	9	5	4	5
4	AND	valence	0,000	24	77	36	41	6	3	5	3	19		16	6	8	4	7
5	IN	valence	0,000	16	67	36	31	9	4	4	9	10		6	5	7	7	6
6	A	valence	0,000	22	54	33	21	3	4	6	12	8			7	7	4	3
7	THAT	valence	0,000	14	40	23	17	6	5	4	7	1			7	6	2	2
8	TO	valence	0,000	9	39	28	11	5	4	8	6	5		2	1	4	3	1
9	IS	valence	0,000	12	39	19	20	4	8	1	4	2		5	3	4	5	3
10	VERB	valence	0,000	8	38	19	19	1	4	7	7			5	1	11	5	2
11	AS	valence	0,000	13	38	19	19	6	4	5	4			5	1	8	5	
12	SUFFIXES	valence	0,000	1	31	12	19	2	1	6	3				13	4		2
13	BY	valence	0,000	6	31	8	23	2	2	1	2	1		11	1	3	4	4
14	7	valence	0,000	2	25	16	9		1	1	14						4	5
15	ARE	valence	0,000	8	24	9	15	2	3	2		2		3	4	1	5	2
16	WITH	valence	0,000	7	22	12	10	3	4	3		2		1	3		3	3
17	CHAPTER	valence	0,000	2	21	18	3	2	1	13	2				1			2
18	OR	valence	0,000	15	21	6	15	2	1			3		8	3	1	2	1
19	S	valence	0,000	3	21	21	0		2		5	14						
20	VERBS	valence	0,000	6	21	8	13	2	4	1		1			3	5	3	2
21	2	valence	0,000	6	20	9	11		4	2		3		2			6	3
22	ONE	valence	0,000	3	20	3	17					3			9	3	2	3
23	3	valence	0,000	6	20	15	5	2	6	2	4	1				2	1	2
24	SPECIFIERS	valence	0,000	1	19	0	19							19				
25	CHANGE	valence	0,000	6	18	11	7			4	7			4	2			1
26	USING	valence	0,000	3	18	2	16				1	1		14		2		
27	PATTERNS	valence	0,000	4	17	1	16	1						15		1		
28	DETERMININ	valence	0,000	1	16	16	0				2	14						
29	SUFFIX	valence	0,000	1	16	10	6	1	2	7					4	2		
30	WHICH	valence	0,000	5	16	11	5	2	2	3	4			1		2		2
31	INCREASING	valence	0,000	2	15	0	15							15				
32	ANALYSIS	valence	0,000	1	15	0	15									14		1
33	SENTIMENT	valence	0,000	1	15	0	15								14		1	
34	SPECIFIER	valence	0,000	1	14	1	13	1						9		2	1	1
35	CHANGING	valence	0,000	3	13	0	13							13				
36	BE	valence	0,000	5	13	6	7	1	2	3					5	2		
37	ROOT	valence	0,000	4	13	9	4		1		1	7					4	

Figure 11. Collocates for the word VALENCE.

As you can realize, the word VALENCE is in the first position, and the total value (345) is in the centre column (in red). Let's pretend we want to analyze the relationship between the words VALENCE and SUFFIXES (line 12); we can see that SUFFIXES appears 13 times (in red) two positions right from VALENCE. If you want to study more this relationship, just click twice in the number 13 and you get back the concordance tab, with these two words chosen (figure 12).



N	Concordance	Word #	Sent #	Sent. Pos	Para #	Para	f	Hei	Sen	f	File	Date
1	100 Table 11. Attested Combinations of Valence-Changing Suffixes.....	520	63	83%	0	2%	0	2%	0	2%	VALENCE CHANGE AND ( 2011/nov/ 7	
2	39 3.6 Summary of Valence Reducing Suffixes.	287	18	75%	0	1%	0	1%	0	1%	VALENCE CHANGE AND ( 2011/nov/ 3	
3	in all of the morphological verb-forming processes. Valence-changing suffixes may combine with each	24.370	1.662	13%	0	85%	0	85%	0	85%	VALENCE CHANGE AND ( 2011/nov/ 1	
4	(m). 100 101 Table 11. Attested Combinations of Valence-Changing Suffixes Derivation Position 1	23.489	1.607	17%	0	82%	0	82%	0	82%	VALENCE CHANGE AND ( 2011/nov/ 1	
5	v □ s □ ? -vis other constituents. 4.4 Summary of Valence Increasing Suffixes I here summarize the	19.497	1.336	31%	0	68%	0	68%	0	68%	VALENCE CHANGE AND ( 2011/nov/ 1	
6	73 4.4 Summary of Valence Increasing Suffixes .	320	22	75%	0	1%	0	1%	0	1%	VALENCE CHANGE AND ( 2011/nov/ 4	
7	10 Table 3. Valence Reducing Suffixes .	456	47	50%	0	2%	0	2%	0	2%	VALENCE CHANGE AND ( 2011/nov/ 7	
8	Verb Roots) ..... 69 Table 6. Valence Increasing Suffixes .	482	53	50%	0	2%	0	2%	0	2%	VALENCE CHANGE AND ( 2011/nov/ 7	
9	the applicative suffix □. 3.6 Summary of Valence Reducing Suffixes The most important	10.432	752	38%	0	36%	0	36%	0	36%	VALENCE CHANGE AND ( 2011/nov/ 7	
10	most important information about Mbonge's five valence-reducing suffixes is summarized in Table 3.	10.442	753	33%	0	36%	0	36%	0	36%	VALENCE CHANGE AND ( 2011/nov/ 7	
11	reasons and with different results. Table 3. Valence Reducing Suffixes Suffix Description	10.468	757	2%	0	37%	0	37%	0	37%	VALENCE CHANGE AND ( 2011/nov/ 7	
12	and with different results. 82 Table 6. Valence Increasing Suffixes Suffix Description	19.536	1.341	7%	0	68%	0	68%	0	68%	VALENCE CHANGE AND ( 2011/nov/ 1	
13	11 shows all the attested suffix combinations of the valence-changing suffixes. Table 10. Attested	23.294	1.588	92%	0	81%	0	81%	0	81%	VALENCE CHANGE AND ( 2011/nov/ 1	
14	282 5.3.1.3. Differences in the valence patterns .	1.959	160	88%	0	2%	0	2%	0	2%	Criteria for the Validation b 2013/mar/ 3	
15	2012) 130 Also, a second table illustrates the valence patterns of the LU, i.e. its combinatorial	40.322	1.796	41%	0	36%	0	36%	0	36%	Criteria for the Validation b 2013/mar/ 3	
16	the FE is further enriched with the specification of valence patterns, the semantic and syntactic	4.118	182	48%	0	50%	0	50%	0	50%	Process-oriented terminolo 2012/mar/ 4	
17	realizations of the different FEs, as well as their valence patterns or mappings between semantic and	3.506	161	77%	0	43%	0	43%	0	43%	Process-oriented terminolo 2012/mar/ 4	
18	issues? RQ1c Can we automatically determine the valence (positive, negative) of these semantic	4.744	187	57%	0	21%	0	21%	0	21%	Semantic Network Analysis 2011/nov/ 7	
19	(who is acting and who is acted upon) and sign (or valence polarity: is the relation positive or	33.053	1.286	77%	1	67%	0	39%	0	39%	Semantic Network Analysis 2011/nov/ 7	
20	and a constructional description of the TAM and valence patterns resolve the apparent	23.376	1.139	70%	0	39%	0	39%	0	39%	Constructional Morphology 2011/nov/ 7	
21	constituents are NP + VP + ADVP. Table 6. Valence patterns of TIDE Table 7. Valence patterns	4.189	186	29%	0	51%	0	51%	0	51%	Process-oriented terminolo 2012/mar/ 5	
22	, resolve1 and suprir1. 5.3.1.3. Differences in the valence patterns About 6% of the total number of	83.552	3.582	10%	0	75%	0	75%	0	75%	Criteria for the Validation b 2013/mar/ 7	
23	of determinar2. The Portuguese verb admits three valence patterns: JUDGE (Sub. NP)	83.838	3.593	70%	0	75%	0	75%	0	75%	Criteria for the Validation b 2013/mar/ 7	
24	structures of the verbs revealed that the verbs' valence patterns are different. For instance,	83.602	3.582	95%	0	75%	0	75%	0	75%	Criteria for the Validation b 2013/mar/ 7	
25	130 Figure 21. Lexical entry report of argue: valence patterns (FrameNet 2012)..... 131	2.853	297	70%	0	3%	0	3%	0	3%	Criteria for the Validation b 2013/mar/ 7	
26	ADVP. Table 6. Valence patterns of TIDE Table 7. Valence patterns of MAREA (1a) TIDE to occur	4.195	187	13%	0	51%	0	51%	0	51%	Process-oriented terminolo 2012/mar/ 5	
27	and prepositions as well as to identify their valence patterns (the ways in which the semantic	39.017	1.726	74%	0	35%	0	35%	0	35%	Criteria for the Validation b 2013/mar/ 7	
28	occur in the contexts of determinar2. Among the valence patterns that the term require1 admits the	83.969	3.613	18%	0	75%	0	75%	0	75%	Criteria for the Validation b 2013/mar/ 7	
29	njn □ ?kd □ qm? . Each suffix indicates an increase in valence, resulting in a net addition of two 104	24.184	1.651	47%	0	84%	0	84%	0	84%	VALENCE CHANGE AND ( 2011/nov/ 1	
30	in section 3.2.1.3. For this reason, the sense and valence rich descriptions provided by FrameNet and	40.783	1.815	23%	0	36%	0	36%	0	36%	Criteria for the Validation b 2013/mar/ 3	
31	a ectedness marking, to a set of voice and valence-related alternations. However, all of these	64.585	2.506	98%	0	86%	0	86%	0	86%	Constructional Morphology 2011/nov/ 7	
32	NP the bottom of the barrel is empty because the valence requirement has been satis ed, 4.1.	57.438	2.004	89%	0	56%	0	56%	0	56%	A CONSTRUCTIONAL API 2011/nov/ 1	
33	Crime Safety Public Order & Safety Crime Justice Valence Rule of Law Economic Growth Economic	56.990	2.231	51%	3	6%	0	68%	0	68%	Semantic Network Analysis 2011/nov/ 1	
34	Person Definite Object) (Agreement= Verb root -Valence -Scenario -Reflexive/ -Negation -N-Object	6.224	374	44%	0	16%	0	16%	0	16%	The Morphology of Modern 2011/nov/ 3	
35	Person Indefinite Object) (Agreement= Verb root -Valence -Scenario -Reflexive/ -Negation -N-Object	6.252	374	73%	0	16%	0	16%	0	16%	The Morphology of Modern 2011/nov/ 3	
36	. V. VI. VII. VIII. IX. X. XI. Agreement= Verb root -Valence -Scenario -Reflex. -Negation -Negation	6.628	404	71%	0	17%	0	17%	0	17%	The Morphology of Modern 2011/nov/ 3	
37	Depend and Depend Depend) (Agreement= Verb root -Valence -Scenario -Negation -N-Object Central	6.200	374	73%	0	16%	0	16%	0	16%	The Morphology of Modern 2011/nov/ 3	

Figure 12. Concordance lines with the words VALENCE and SUFFIXES

Just with this VALENCE example, we can study hundreds of combinations. For example, if you analyze picture 6, you can notice what are the prepositions that follow VALENCE one position to the right: of (30 times), in (6 times), by (11 times), etc. From this analysis, you could postulate that the most common preposition that follows the word VALENCE is of. A very simple postulation, but based on facts.

### Clusters/N-Grams

According to Scott (2012), clusters are:

[...] words which are found repeatedly together in each others' company, in sequence. They represent a tighter relationship than collocates, more like multi-word units or groups or phrases. (I call them clusters because groups and phrases already have uses in grammar and because simply being found together in software doesn't guarantee they are true multi-word units.) Biber calls them "lexical bundles".

Language is phrasal and textual. It is not helpful to see it as a matter of selecting a word to fill a grammatical "slot" as implied by structural theories. Words keep company: the extreme example is idiom where they're bound tightly to each other, but all words have a tendency to cluster together with some others. These clustering relations may involve colligation (e.g. the relationship between depend and on), collocation, and semantic prosody (the tendency for cause to come with negative effects such as accident, trouble, etc.).

Following the same example above, let's see the clusters (from 3 to 5 words, with a minimum of 5 examples) involving VALENCE (Figure 13):

N	Cluster	Freq.	Length	Related
1	THE VALENCE OF THE	23	3	THE VALENCE OF THE (11), THE VALENCE OF THE VERB (6)
2	USING SENTIMENT ANALYSIS	15	3	VALENCE USING SENTIMENT ANALYSIS (14), DETERMINING VALENCE USING SENTIMENT ANA
3	DETERMINING VALENCE USING SENTIMENT	14	4	DETERMINING VALENCE USING (14), VALENCE USING SENTIMENT (14), DETERMINING VALENCE
4	DETERMINING VALENCE USING	14	3	DETERMINING VALENCE USING SENTIMENT (14), DETERMINING VALENCE USING SENTIMENT
5	CHAPTER 7 DETERMINING	14	3	CHAPTER 7 DETERMINING VALENCE (13), CHAPTER 7 DETERMINING VALENCE USING (13)
6	VALENCE USING SENTIMENT	14	3	DETERMINING VALENCE USING SENTIMENT (14), VALENCE USING SENTIMENT ANALYSIS (14)
7	VALENCE USING SENTIMENT ANALYSIS	14	4	USING SENTIMENT ANALYSIS (15), VALENCE USING SENTIMENT (14), DETERMINING VALENCE
8	DETERMINING VALENCE USING SENTIMENT ANALYSIS	14	5	USING SENTIMENT ANALYSIS (15), DETERMINING VALENCE USING SENTIMENT (14), DETERMI
9	VERB S VALENCE	14	3	A VERB S VALENCE (6), THE VERB S VALENCE (6)
10	7 DETERMINING VALENCE	14	3	7 DETERMINING VALENCE USING (14), 7 DETERMINING VALENCE USING SENTIMENT (14), CHA
11	7 DETERMINING VALENCE USING	14	4	DETERMINING VALENCE USING (14), 7 DETERMINING VALENCE (14), 7 DETERMINING VALENCE
12	7 DETERMINING VALENCE USING SENTIMENT	14	5	DETERMINING VALENCE USING SENTIMENT (14), DETERMINING VALENCE USING (14), 7 DETE
13	CHAPTER 7 DETERMINING VALENCE	13	4	CHAPTER 7 DETERMINING (14), 7 DETERMINING VALENCE (14), CHAPTER 7 DETERMINING VAI
14	CHAPTER 7 DETERMINING VALENCE USING	13	5	DETERMINING VALENCE USING (14), CHAPTER 7 DETERMINING (14), 7 DETERMINING VALENCE
15	VALENCE OF THE	12	3	THE VALENCE OF THE (11), VALENCE OF THE VERB (6), THE VALENCE OF THE VERB (6)
16	THE VALENCE OF THE	11	4	THE VALENCE OF (23), VALENCE OF THE (12), THE VALENCE OF THE VERB (6)
17	A VERB S	10	3	TO A VERB S (6), A VERB S VALENCE (6)
18	AS WELL AS	9	3	
19	VALENCE BY ONE	9	3	VALENCE BY ONE AND (5)
20	VERB ROOT VALENCE	8	3	VERB ROOT VALENCE SCENARIO (6), AGREEMENT VERB ROOT VALENCE (5)
21	TO A VERB	7	3	TO A VERB S (6)
22	OF THE VALENCE	7	3	
23	IN THE VALENCE	7	3	
24	TO A VERB S	6	4	A VERB S (10), TO A VERB (7)
25	VALENCE OF THE VERB	6	4	VALENCE OF THE (12), OF THE VERB (6), THE VALENCE OF THE VERB (6)
26	VALENCE OF A	6	3	
27	VALENCE REDUCING SUFFIXES	6	3	
28	OF THE VERB	6	3	VALENCE OF THE VERB (6), THE VALENCE OF THE VERB (6)
29	A VERB S VALENCE	6	4	VERB S VALENCE (14), A VERB S (10)
30	VERB ROOT VALENCE SCENARIO	6	4	VERB ROOT VALENCE (6), ROOT VALENCE SCENARIO (6)
31	ROOT VALENCE SCENARIO	6	3	VERB ROOT VALENCE SCENARIO (6)
32	THE VERB S VALENCE	6	4	VERB S VALENCE (14), THE VERB S (6)
33	THE VERB S	6	3	THE VERB S VALENCE (6)
34	THE VALENCE OF THE VERB	6	5	THE VALENCE OF (23), VALENCE OF THE (12), THE VALENCE OF THE (11), VALENCE OF THE V
35	BY ONE AND	5	3	VALENCE BY ONE AND (5)
36	AGREEMENT VERB ROOT VALENCE	5	4	VERB ROOT VALENCE (8), AGREEMENT VERB ROOT (5)
37	AGREEMENT VERB ROOT	5	3	AGREEMENT VERB ROOT VALENCE (5)

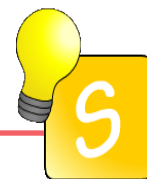
Figure 13. Cluster tab for the word VALENCE

In the example 20, we have the cluster verb<sup>7</sup> root valence, with 8 examples. If you want to analyze this cluster, just go back to the collocates tab (take a look at line 37 of figure 11) and choose root. The result is the following (figure 14):

N	Concordance	Word	Date	%	
1	. IV. V. VI. VII. VIII. IX. X. XI. Agreement= Verb root -Valence -Scenario -Reflex. -Valence -Negation N-Object	6.82 4 7 0 11	0 11	The Morpl 2011/nov	37%
2	Person and Second Person) Agreement= Verb root -Valence -Scenario -Negation -N-Object -Central -Tense	6.20 3 1 2 0 11	0 11	The Morpl 2011/nov	35%
3	(Third Person Definite Object) (Agreement= Verb root -Valence -Scenario -Reflexive/ -Negation -N-Object	6.22 3 1 4 0 11	0 11	The Morpl 2011/nov	35%
4	Positions I. II. III. IV. V. VI. VII. VIII. IX. X. Verb root -Valence -Scenario Animacy -Central Participant/	6.88 4 6 0 11	0 11	The Morpl 2011/nov	38%
5	Verb Structure Animate Subject (Agreement Verb root -Valence -Negation -N-Object -Person -Tense Clitic)=	5.94 3 2 0 11	0 11	The Morpl 2011/nov	34%
6	Marker Agreement Inanimate Subject Verb root -Valence -Negation -Tense Specifier 27 Animate	5.95 3 5 0 11	0 11	The Morpl 2011/nov	34%
7	(Third Person Indefinite Object) (Agreement= Verb root -Valence -Scenario -Reflexive/ -Negation -N-Object	6.25 3 1 7 0 11	0 11	The Morpl 2011/nov	35%
8	entities, and 4) display no significant differences in their valence patterns. Equivalents were considered partial	84.8 3 9 0 71	0 71	Criteria fo 2013/mai	74%
9	(Reciprocity). Figure 21. Lexical entry report of argue: valence patterns (FrameNet 2012) 132 The Using and	40.4 1 2 0 31	0 31	Criteria fo 2013/mai	37%
10	2012) 130 Also, a second table illustrates the valence patterns of the LU, i.e. its combinatorial	40.3 1 4 0 31	0 31	Criteria fo 2013/mai	37%
11	282 5.3.1.3. Differences in the valence patterns .....	1.95 1 6 0 21	0 21	Criteria fo 2013/mai	3%
12	130 Figure 21. Lexical entry report of argue: valence patterns (FrameNet 2012) .....	2.85 2 7 0 31	0 31	Criteria fo 2013/mai	5%
13	constituents are NP + VP + ADVP. Table 6. Valence patterns of TIDE Table 7. Valence patterns of	4.18 1 1 2 0 51	0 51	Process-c 2012/mai	50%
14	(who is acting and who is acted upon) and sign (or valence, polarity: is the relation positive or negative?) or	33.0 1 7 1 91	0 31	Semantic 2011/nov	76%
15	+ ADVP. Table 6. Valence patterns of TIDE Table 7. Valence patterns of MAREA (1a) TIDE to occur	4.19 1 1 1 0 51	0 51	Process-c 2012/mai	50%
16	of the FE is further enriched with the specification of valence patterns, the semantic and syntactic templates	4.11 1 1 4 0 51	0 51	Process-c 2012/mai	49%
17	never occur in the contexts of determinar2. Among the valence patterns that the term require1 admits the only	83.9 3 1 1 0 71	0 71	Criteria fo 2013/mai	74%
18	passive sentences, can also be stated in terms of the valence or dtrs attributes. This is discussed in Section	70.8 2 9 0 71	0 71	A CONST 2011/nov	135%
19	actantial structures of the verbs revealed that the verbs' valence patterns are different. For instance, commit2,	83.6 3 9 0 71	0 71	Criteria fo 2013/mai	73%
20	. This chapter further enriches the relations with valence or polarity, creating a directed and signed	51.4 1 4 1 2 81	0 61	Semantic 2011/nov	119%
21	meanings specific of this field as well as a syntactic valence or combinatory value. The concentration of such	1.13 5 9 0 21	0 21	SEMANTI 2012/mai	21%
22	meanings specific of this field as well as a syntactic valence or combinatory value. Naturally, such noun	476 2 9 0 11	0 11	Frame-Ba 2012/mai	15%
23	lexical and a constructional description of the TAM and valence patterns resolve the apparent inconsistencies of	29.3 1 7 0 31	0 31	Constructi 2011/nov	73%
24	realizations of the different FEs, as well as their valence patterns or mappings between semantic and	3.50 1 6 7 0 41	0 41	Process-c 2012/mai	42%
25	, adverbs, and prepositions as well as to identify their valence patterns (the ways in which the semantic	39.0 1 7 0 31	0 31	Criteria fo 2013/mai	36%
26	of determinar2. The Portuguese verb admits three valence patterns: JUDGE (Sub. NP) PROTAGONIST	83.8 3 7 0 71	0 71	Criteria fo 2013/mai	73%
27	, resolve1 and suprir1. 5.3.1.3. Differences in the valence patterns About 6% of the total number of	83.5 3 1 1 0 71	0 71	Criteria fo 2013/mai	73%
28	and issues? RQ1c Can we automatically determine the valence (positive, negative) of these semantic relations?	4.74 1 1 5 0 21	0 61	Semantic 2011/nov	9%
29	construction to translate this phrase although there is no valence reduction apparent in the Abenaki. Leavitt	11.9 7 7 0 31	0 31	The Morpl 2011/nov	63%
30	: he'elim In languages that morphologically mark valence reduction, these unaccusatives often bear	2.91 1 6 0 21	0 21	Hidden en 2013/mar	26%
31	discusses the applicative suffix ,d. 3.6 Summary of Valence Reducing Suffixes The most important	10.4 7 3 0 31	0 31	VALENCE 2011/nov	74%
32	verbs do not appear in a morphological form typical of valence reducing operations: (14) ha-ra'ayon xamak	3.02 1 3 0 21	0 21	Hidden en 2013/mar	27%
33	for different reasons and with different results. Table 3. Valence Reducing Suffixes Suffix Description Function	10.4 7 2 0 31	0 31	VALENCE 2011/nov	74%
34	in section 3.2.1.3. For this reason, the sense and valence rich descriptions provided by FrameNet and	40.7 1 2 0 31	0 31	Criteria fo 2013/mai	37%
35	Crime Safety Public Order & Safety Crime Justice Valence Rule of Law Economic Growth Economic	56.9 2 5 1 3 61	0 61	Semantic 2011/nov	133%
36	njn. Each suffix indicates an increase in valence, resulting in a net addition of two 104 arguments	24.1 1 4 0 81	0 81	VALENCE 2011/nov	168%
37	participant a etchedness marking, to a set of voice and valence-related alternations. However, all of these	64.5 2 9 0 81	0 81	Constructi 2011/nov	169%

Figure 14. Concordance lines for the cluster verb root valence.

Now you can analyze this cluster according to your research.



Although WST is a paid program, the author (Mike Scott) makes the version 5 available for free in his site: <https://lexically.net/wordsmith/>

<sup>7</sup> If we would use the n-gram concept, this would be a trigram.



# MÓDULO 2

## Corpus Linguistics: Teaching and Learning

### **Basic Contents**

- At school, at home, self-learning
- Online corpora

### **Objectives**

- Realize that Corpus Linguistics is a useful tool for teachers and learners.
- Access a lot of free tools available on the Internet.

## CORPUS LINGUISTICS: TEACHING AND LEARNING

### ACTIVITY 5



Video class, module 2. Watch the professor's hints about the subjects that are going to be worked in this module.

### At school, at home, self-learning

Corpus Linguistics can be a very powerful teaching and learning tool. As a teacher, you can show your students how to solve their linguistic problems by themselves. As a learner, you can learn a lot of language patterns and also solve your doubts. Differently from other areas in Linguistics, Corpus Linguistics tools and texts are generally available at free cost on the Internet.



The book **Corpora no ensino de línguas estrangeiras** brings many ideas how corpora can be worked to teach. Read a review of this book in here:

<http://www.seer.ufu.br/index.php/dominiosdelinguagem/article/view/12335>

We're going to present, in this module, some of the tools you can freely use on the Internet and how to interact with them.

### COCA

The Corpus of Contemporary American English is the biggest freely available corpus on the Internet. Nowadays, with 450 million words (it gets more 25 million words every year), it presents a very easy and intuitive interface. From there, you can also access other corpora using the same search structure.

Something very important about COCA is the balance among its texts.

You can find its portal here: <https://www.english-corpora.org/coca/>. Register yourself,



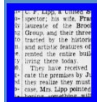
Figure 15. Registration corner. Source: COCA.

At the first page of COCA, there are links for many other sub corpora, specific features of the system and downloads you can try:



Figure 16. COCA's upper menu.

Table 2. Features related to COCA.

<p>Word and Phrase (analyze texts)</p>  <p>Click this icon</p>	<p>Enter entire texts and see detailed frequency information on the words in the text and create word lists based on your text. Click through the words to see detailed information on any word. Highlight phrases in your text and have it search for related phrases in COCA. Search and browse the most complete frequency dictionary of English. See detailed information (all on one page) -- definition, frequency by genre, collocates (nearby words), concordance lines, synonyms, and Wordnet-related words, all with useful links from one resource to another.</p>
<p><a href="#">Word Frequency</a></p>	<p>You can also download lists showing the frequency of the top 60,000 lemmas by genre (and sub-genre), as well as the top 200-300 collocates (nearby words) for these lemmas (4,800,000 node/collocate pairs). There is also a free list of the top 5,000 lemmas in COCA. And now you can download the 100,000 integrated word list from COCA, COHA, BNC, and SOAP -- the largest, corrected frequency list of English.</p>
<p><a href="#">Collocates</a></p>	<p>Download lists with the top 200-300 collocates (nearby words) for 60,000 different lemmas -- 4,300,000 node/collocate pairs in all.</p>
<p><a href="#">N-grams</a></p>	<p>Download free lists containing the top 1,000,000 2-grams (two word sequences), 3-grams, 4-grams, and 5-grams in COCA. There are also other lists that contain the frequency of all 2, 3, and 4-grams (up to 155 million rows of data).</p>
<p><a href="#">Academic vocabulary</a></p>	<p>Download free lists containing "core" academic words in 120 million words of COCA-Academic texts (including grouping by word families), as well as the top 20,000 words overall in COCA-Academic. See <i>Applied Linguistics</i> article, or compare to the Academic Word List (Coxhead, 2000).</p>
<p><a href="#">Word and Phrase (academic)</a></p>	<p>Similar to the two Word and Phrase resources below, but limited strictly to the 120 million words of academic texts in COCA. Get detailed information on words and phrases, frequency by sub-genre (e.g. Law, Medicine, Science, Business, Humanities), and concordances and collocates in just the academic text. Also, analyze entire academic texts.</p>

On the left upper left site, you find the search display to star your queries (Figure 17).

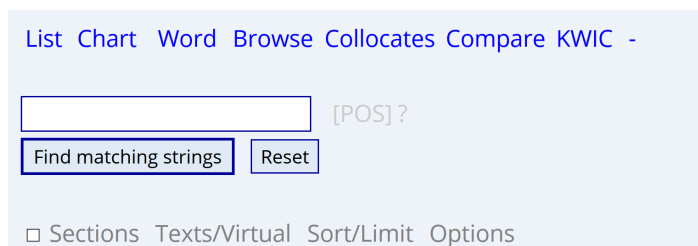


Figure 17. COCA's search display.

Using this search display, let's try a basic query? You're producing a text and suddenly a doubt comes out: we use at the Internet, in the Internet or on the Internet? You know this is a kind of problem the grammar books are not going to solve. Instead of looking for a native speaker, you can search the three possibilities using the COCA. Click on LIST, write down the possibility you want to analyze (for example, at the Internet) and click the FIND MATCHING STRINGS button. What you get (figure 18) is a screen on the right side, with the frequency found (in this case, 155 entries); note that the middle menu changed from the first to the second option (SEARCH > FREQUENCY).



Figure 18. COCA's query result.

If you click on the combination you are searching (at the Internet), you get a KWIC screen with all the concordances (figure 19); again the middle menu changed, now from he second to the third option (FREQUENCY > CONTEXT):

Corpus of Contemporary American English

SEARCH FREQUENCY CONTEXT ACCOUNT

FIND SAMPLE: 100  
PAGE: << < 1 / 2 > >>

CLICK FOR MORE CONTEXT  [?] SAVE LIST CHOOSE LIST CREATE NEW LIST [?]

Line	Year	Genre	Source	Concordance
1	2019	FIC	Obsidian	A B C memory. It was one of the many words he wanted to research at the Internet caf when he returned home. # The Black girl talked about the same
2	2018	SPOK	NPR_Sunday	A B C thing where they're either like super precocious because they're so good at the Internet and they're like YouTube billionaires or they're hopeless
3	2018	MAG	Ars Technica	A B C Bioglio et al. applied their algorithm to information collected on 47,000 films at the Internet Movie Database (IMDB). (The authors caution that the
4	2017	SPOK	NPR_FreshAir	A B C a lot of people don't like him. And when I looked at the internet, I discovered there is this bunch of people that don't like you.
5	2017	SPOK	CNN: Primetime Justice	A B C type of privacy issues. You have -- so many young people look at the Internet as, This is my private thing, and they don't think about everyone
6	2017	MAG	A.V. Club	A B C this sense it is perhaps the perfect thing to listen to while looking at the internet today. Don't say we didn't warn you.
7	2017	MOV	After Porn Ends 2	A B C and then it destroyed the business pretty much. People can just look at the Internet and get whatever they want for free. There's no need to put a
8	2016	MAG	TechCrunch	A B C New rules in the era of " things " # Let's look at the Internet of Things. This is the hardware industry of the future, and it will
9	2016	MAG	Fortune	A B C competitors, Verizon Communications vz and AT&T t, are also taking aim at the Internet of things opportunity. The bigger players are aiming to es
10	2016	NEWS	Washington Post	A B C Drudge was feeling celebratory on Twitter. # (Drudge, no dummy at the Internet, purges his tweets regularly, which is why we've included a screen
11	2015	SPOK	ABC: The View	A B C internet companies. Is there a responsibility for someone to take a look at the internet and look at it for what it is, which is a tool for spreading
12	2015	MAG	NatGeog	A B C York City. Adeoti makes twice the salary he made as the manager at the Internet cafe. But all this exposure to money and movies had whetted his
13	2015	ACAD	QuartRevDistanceEd	A B C : What do students think, want, and do? Paper presented at the Internet Research 14.0, Denver, CO. # Dennen, V. P., & Burner,
14	2015	ACAD	QuartRevDistanceEd	A B C and the higher education classroom: Student preferences and attitudes. Paper presented at the Internet Research 15.0, Daegu, South Korea. # Di
15	2015	TV	The Amazing Race	A B C are the Olympians. - We were going downstairs... - To look at the Internet. - And we randomly ran across... - The Olympians. We're going
16	2014	SPOK	PBS_Newshour	A B C 're really more volatile because of the expectation, particularly when you look at the Internet stocks, then sort of new technology, social media stc
17	2014	FIC	Ploughshares	A B C chances of remission were excellent, everyone said. " Don't look at the internet, " her new oncologist warned her, and Merel, numb and childlike v
18	2014	FIC	WarLitArts	A B C been contracted out from elsewhere. The guards at the chow hall and at the internet cafe were Ugandan Army, many of whom had seen action di
19	2014	MAG	PopMech	A B C Past eras simply can't compete. # For proof, just look at the Internet. Nobody could have imagined typing something and everyone in the world b
20	2013	SPOK	CNN: CNN Live Event	A B C . So certainly, because there are around two million citizens who look at the internet, the government is allowing this discussion to happen. So, it c

Figure 19. COCA's concordance lines for the query at the Internet

As you can see, the platform gives you a lot of information about each line of concordance. In line 1, for example, we have: the year the material was published (2019), the media it was published (fiction), the name of the media (Obsidian). If you click on this links, you get part of the text with the information highlighted (figure 20); once more the middle menu changed, this time from the third to the fourth option (CONTEXT > CONTEXT+):

Corpus of Contemporary American English

SEARCH FREQUENCY CONTEXT CONTEXT +

Source information:

Source	FIC: Obsidian: Literature in the African Diaspora
Date	2019 (2017/03/22)
Publication information	Vol. 43, Issue 1
Title	PLASTIC CITY
Author	Leslie Ann Murray

Expanded context:

the Black girl's shoulders until she moaned and then worked on her lower back. He moved his hands around her lower back until she winced, " Ouch. " He learned that " ouch " meant it was the spot that most needed massaging. # " Men suck, " she said. " I wish I was not committed to heterosexuality. " # " Heterosexuality. " He did not have his notebook, so he quietly repeated the word until it lodged into his memory. It was one of the many words he wanted to research at the Internet caf when he returned home. # The Black girl talked about the same things all tourists talked about -- romantic relationships, family life, work life and life. Her present stress was Marius-related. Whenever tourists started talking about this personal side of their lives, he blocked them out and searched for his Zen. # " What do you think? " The Black girl asked. " Are you listening to me? " # " Um, love should not be so complicated, " he said.

Source information:

Date	2012
Publication information	Jan/Feb 2012
Title	Net Worth
Author	Negulescu, Kris Carpenter
Source	technology Review

Expanded context:

preservation. Is this too far-out to imagine? Perhaps. But such cooperation is appearing within international research communities and cultural groups in both Europe and the United States. This work creates a foundation we can build upon. Only by encouraging this type of collaboration among like minded communities can we hope to preserve any significant slice of the Web. The future does not afford anyone the luxury of the unlimited time, funds, computing power, and storage capacity that would be needed to do it alone. AuthorAffiliation # KRIS CARPENTER NEGULESCU IS DIRECTOR OF WEB ARCHIVING AT THE INTERNET ARCHIVE, A NONPROFIT INTERNET LIBRARY THAT PRESERVES DIGITAL CONTENT 928063660 INFORMATION TECHNOLOGY # Touch screens that work through fabric # SOURCE: \* POKETTTOUCH: THROUGH-FABRIC CAPACITIVE TOUCH INPUT \* # T. Scott Saponas et al. # Proceedings of the 24th ACM Symposium on User Interface Software and Technology, Santa Barbara, California, October 10-19, 2011 # RESULTS: Researchers at Microsoft created a touch screen that can be operated

Figure 20. Expanded concordance line for the first example.

## ACTIVITY 6 - SEARCH



Using the COCA, search the other possibilities given above using a *preposition + the Internet*. What are your findings? Can you systematize a rule of usage?



## ACTIVITY 7 - DISCUSSION



Discuss with the other students, using the AVA forum, what is the best way to use and the differences among the structure: *preposition + the Internet* given above. The participation in the discussion will be evaluated.

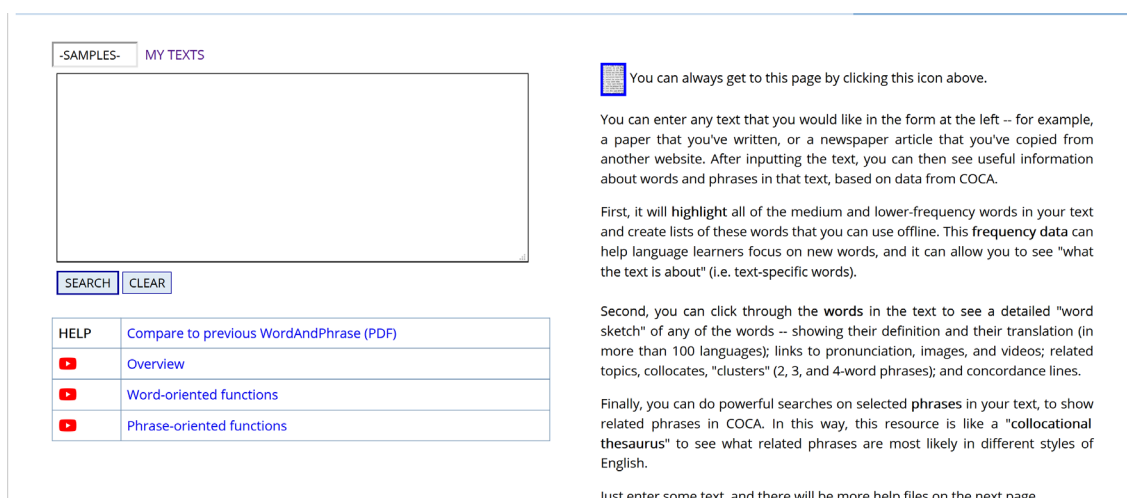
The activities we performed before are just one among a lot of others we can try using the COCA platform. Take a look at possible searches using combinations and wildcards in table 3:



Table 3. Query possibilities using the COCA interface.

Type of search	COCA-General
Specific word or phrase	I guess
Substring	*al_j
Lemma (forms of a word)	CONJ PRON BE like , ( and she was like , )
Part of speech	ADJ eyes
Synonyms	=strong
User-defined lists	@colors @clothes
Sortable concordance lines	fathom
Collocates (nearby words)	BREAK_v
-- Use Mutual Information score	BREAK_v
-- Compare two words	utter / sheer

Let's try, now, another important feature of COCA: the Word and Phrase tool (figure 21). Access using the  icon.



-SAMPLES- MY TEXTS

SEARCH CLEAR

HELP

- Compare to previous WordAndPhrase (PDF)
- Overview
- Word-oriented functions
- Phrase-oriented functions

You can always get to this page by clicking this icon above.

You can enter any text that you would like in the form at the left -- for example, a paper that you've written, or a newspaper article that you've copied from another website. After inputting the text, you can then see useful information about words and phrases in that text, based on data from COCA.

First, it will **highlight** all of the medium and lower-frequency words in your text and create lists of these words that you can use offline. This **frequency data** can help language learners focus on new words, and it can allow you to see "what the text is about" (i.e. text-specific words).

Second, you can click through the **words** in the text to see a detailed "word sketch" of any of the words -- showing their definition and their translation (in more than 100 languages); links to pronunciation, images, and videos; related topics, collocates, "clusters" (2, 3, and 4-word phrases); and concordance lines.

Finally, you can do powerful searches on selected **phrases** in your text, to show related phrases in COCA. In this way, this resource is like a "**collocational thesaurus**" to see what related phrases are most likely in different styles of English.

Just enter some text, and there will be more help files on the next page.

Figure 21. Word and Phrase main page. COCA platform.

Let's try a simple word first. LINGUISTICS is the word we are going to search. Just write the word and click the SEARCH button.

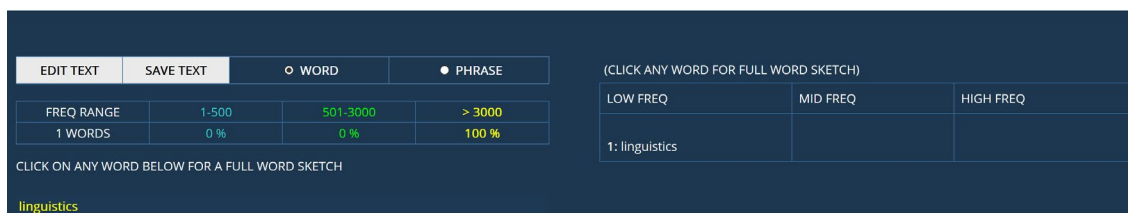


Figure 22. Frequency query for Linguistics.

What do you get in this screen? On the left side, our search word (Linguistics) is classified, according to its frequency, among three levels of words in the COCA corpus: range 1-500 (in blue, extremely common words), range 501-3000 (in green, common words) and range above 3000 (in yellow, rare words or words belonging to specific domains). As you can see, LINGUISTICS belongs to the third range, i.e.<sup>8</sup>, it's not a very common word in English (it doesn't have a very high frequency in the corpus). Now, let's learn more about it: click on the word linguistics (in yellow, left corner). What we get is a complete description of this word in the English language (figure 23). Click on the innumerable links to learn more about it.

**linguistics** (NOUN) #15174

1. the scientific study of language 2. the humanistic study of language and literature

**SYNONYMS** (more): dialectology, grammar, linguistics, morphology, phonology, semantics, syntax

**COLLOCATES** (more):  
 NOUN: professor, university, language, study, association, corpus, department, field  
 VERB: study, teach, combine, reconstruct, specialize, invent, portray, practice  
**ADJ**: applied, computational, cognitive, historical, clinical, modern, literary, arabic  
 ADV: eg, primarily, surely, genetically, per, se, the, flat

**CLUSTERS** (more)

linguistics •	linguistics at • linguistics professor • linguistics in • linguistics department • linguistics from • linguistics to • linguistics with • linguistics has
• linguistics	in linguistics • applied linguistics • computational linguistics • corpus linguistics • historical linguistics • cognitive linguistics • clinical linguistics • to linguistics
linguistics ••	linguistics professor at • linguistics and education • linguistics and literature • linguistics and applied • linguistics and literacies • linguistics and philosophy • linguistics and literary • linguistics and phonetics
•• linguistics	professor of linguistics • for computational linguistics • language and linguistics • degree in linguistics • field of linguistics • in applied linguistics • institute of linguistics • department of linguistics
linguistics •••	linguistics at the university • linguistics and applied language • linguistics and the novel • linguistics and the school • linguistics and evolutionary biology • linguistics at hebrew university • linguistics and technology program • linguistics and literary studies
••• linguistics	association for computational linguistics • summer institute of linguistics • center for applied linguistics • review of applied linguistics • ridiculous interests include linguistics • cognitive science and linguistics • holds appointments in linguistics • course in general linguistics

<sup>8</sup> Learn about Latin abbreviations in English: [https://en.wikipedia.org/wiki/List\\_of\\_Latin\\_abbreviations](https://en.wikipedia.org/wiki/List_of_Latin_abbreviations)

TEXTS / VIRTUAL CORPORA (more)

ACAD:Style • ACAD:Harvard J Law Public Policy • ACAD:Style • ACAD:MC Bioinformatics • WEB:mesa.arizona.edu • WEB:dailywritingtips.com • ACAD:AfricanHist • ACAD:The European Journal of Applied Linguistics and TEFL • ACAD:AI Magazine • BLOG:..agelog.ldc.upenn.edu • ACAD:Style • ACAD:ArabStudies • ACAD:CommCollegeR • ACAD:Laboratory Phonology; Journal of the Association for Laboratory Phonology • WEB:rucss.rutgers.edu • WEB:niemanstoryboard.org • ACAD:Style • ACAD:Style • ACAD:Style • WEB:...enotes.blogspot.com • BLOG:blogs.suntimes.com • WEB:tabletmag.com • BLOG:reason.com • ACAD:Am Folklore •

CONCORDANCE LINES (more)

1	ACAD: 2017: AI Magazine	Equations . Transactions of the Association for Linguistics	2	[ 585-597 ]	# Koncel-Kedziorski , R. ; Roy , S. ;
2	ACAD: 1995: AcademicQs	than on religion , economics , sociology , or language and linguistics.	5	[ # Will	someone inform the professorate that , at a time
3	ACAD: 1996: CurrentPsych	subjects their message would only be used in a linguistics analysis	[ Second	,	in the audience condition subjects gave the
4	ACAD: 1994: AmerStudies	; and the non-membership ERIC Clearinghouse on Language and Linguistics	and Center for	Applied Linguistics	serve as a pool of research
5	ACAD: 1992: AmerStudies	American students.72 # Staff and students of linguistics	and education		as well as other disciplines , have not been
6	ACAD: 2017: AI Magazine	first attempts to put together researchers from linguistics	and from	argumentation	theory .
7	ACAD: 1998: SocialStudies	-- archaeology is one of four subfields , which also include linguistics	and	and	cultural anthropology . Despite its
8	SPOK: 2009: NPR_TalkNation	tactics here . DAVID : Well , we're also talking linguistics	and	pro-life	is - pro-life is anti-choice . Pro-life is - back
9	MAG: 1990: AmSpect	here . # We are not told either whether he studied linguistics	at a	University	or whether it is just his avocation . It
10	NEWS: 2012: NYTimes	Think Jen Johnson 's keypad thumbs . A graduate student in linguistics	at Georgetown	University	, Ms. Johnson , 21 , stumbled onto
11	NEWS: 2012: NYTimes	, Ms. Johnson , 21 , stumbled onto Siletz while studying linguistics	at Swarthmore	College	, which has helped the tribe build its
12	NEWS: 2012: NYTimes	, " said K. David Harrison , an linguistics	at Swarthmore	who	worked with the Siletz tribe and the other
13	WEB: 2012: amazon.com	theories of morality should explain moral behavior , much as linguistics	attempts to	explain	human verbal communication . Philosophers
14	BLOG: 2012: vickiarcher.com	decades of formally studying and/or teaching literature & linguistics	before		realized that my East Tennessee grandmother was
15	BLOG: 2012: reason.com	but as a linguist I need to point out that all linguistics	can tell	us	is that the oldest languages we know anything at
16	ACAD: 2005: AnthropolQ	social distance , rather than linguistics	categories		for example , the Greenberg African language
17	WEB: 2012: ...tword.dictionary....	phoneme (I think -- it 's been a while since linguistics	class		in the English language is " er " - #
18	FIG: 2014: VirginiaQRev	n't have to work for a living . Her degree in linguistics	combined	with	her halting English did n't leave her many options
19	BLOG: 2012: a-sense-of-place.com	lab , information and decision systems lab , and its linguistics	department		opened in 2004 after a lot of funding from
20	NEWS: 1990: WashPost	Worcester , Mass . He received a master 's degree in linguistics	from Georgetown	University	and a doctorate in anthropology
21	ACAD: 2001: AnthropolQ	also how it is interpreted . Recent work in anthropology and linguistics	has	greatly	improved our analysis of interpretation by showing
22	WEB: 2012: deirdremcloskey.com	the cheap talk paradox " ). A linguistics	illustrates	the	point . A very pompous linguist was giving a talk
23	ACAD: 1996: AmerStudies	its function ; # c) insufficient development of linguistics	in the	aspect	verging on culture studies ; # d) the
24	WEB: 2012: amazon.com	perspective of Noam Chomsky 's theories that transformed linguistics	in the	1960s	. Stephen Pinker 's " The Language Instinct "
25	ACAD: 1995: ArtBulletin	the disciplinary and professional structure of knowledge ? Linguistics	is a	discipline	; English is a department ; cinema studies is
26	NEWS: 2013: NYTimes	Hamdallah was educated in the West -- his Ph.D . in linguistics	is from	the	University of Lancaster -- and is not officially tied

Figure 23. Word LINGUISTICS screen.

As you can see in figure 23, the concordance lines present you words with different colors > they represent different word classes. We know that Linguistics is a noun. Can you guess the association among the other colors and word classes?



There are innumerable ways to analyze combinations with Linguistics. Click on the other possibilities (as in figure 24) to get new screens:

See in iWeb Collocates Clusters Topics Dictionary Texts KWIC HELP

Figure 24. Menu to get to new screen

Try also to search with a text, like in figure 25. The result (figure 26) can be analyzed word by word or as a whole.

-SAMPLES- MY TEXTS

America is rich in small businesses. These enterprises account for over 30 million U.S. businesses and some two-thirds of net new jobs. While venture-backed startups generally skew white, male and coastal, these Main Street companies actually look like—and drive—America. To shine a light on these entrepreneurial heroes, Forbes has created the Next 1000. This year-round initiative showcases the ambitious sole proprietors, self-funded shops and pre-revenue startups in every region of the country—all with under \$10 million in revenue or funding and infinite drive and hustle. Fueled by your nominations and screened by top business minds and entrepreneurial superstars, these new faces will number 1,000 by

SEARCH CLEAR

Figure 25. Text search

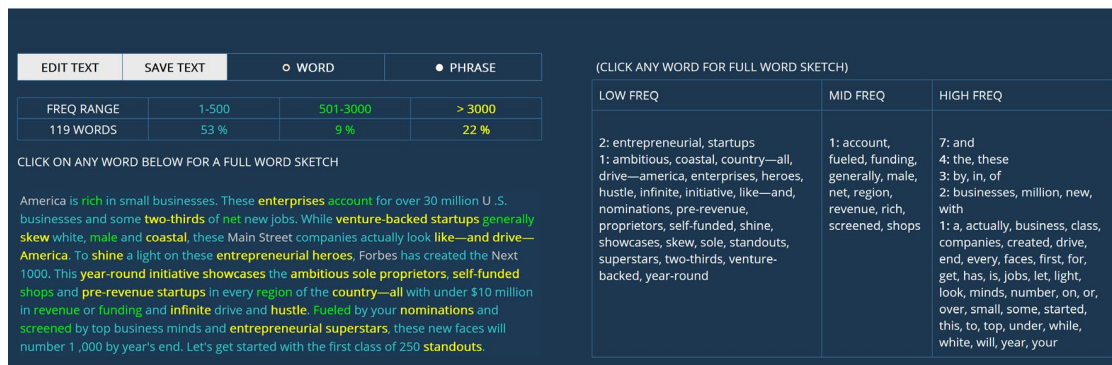


Figure 26. Text analysis

Something very important about COCA: they also have Portuguese corpora, using the same interface of the English corpora. Take a look here (figure 27):

<https://www.corpusdoportugues.org/xp.asp>

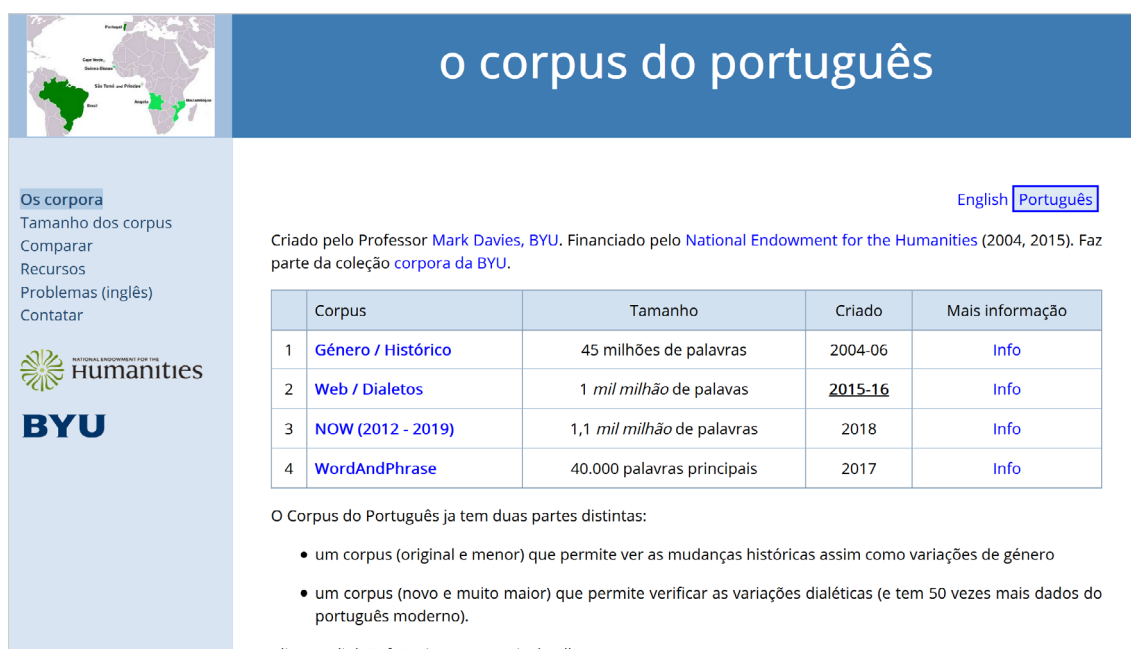


Figure 27. Corpus do Português, COCA platform.



## ACTIVITY 8 - PIPE

It's your time to evaluate a site that works with corpora. What's the idea? You find a site that has tools to work with corpora already compiled (like Projeto COMET, BNC, other COCA platforms, Linguateca, LAEL, Webcorp, etc.) and you're going to describe the site, as we have done with COCA above. You can also print the screens if you want. This work will be evaluated by your tutor.

# MÓDULO 3

## Lexical Analysis Tools

### **Basic Contents**

- Lexical analysis software overview.
- Corpus compilation.
- AntConc suite for corpus analysis.

### **Objectives**

- Have the basic knowledge about how to compile a corpus.
- Use the basic tools (wordlist, keywords and concordance) of a lexical analysis software.

## LEXICAL ANALYSIS TOOLS

### ACTIVITY 9



Video class, module 3. Watch the professor's hints about the subjects that are going to be worked in this module.

There are many software you can use to describe language. The most modern and complete ones must be bought (like WordSmith Tools) or subscribed (like the Sketch Engine).



Although old, the following text shows some differences among them: <http://www.ileel.ufu.br/guifromm/upload/ferramentasdeanaliselexicalcomputadorizadas.pdf>. For a more updated text, comparing WordSmith Tools and Sketch Engine, try this one: <http://www.periodicos.letras.ufmg.br/index.php/relin/article/view/15766>

An example of a very powerful corpus analysis suite is the WordSmith Tools (<http://lexically.net/wordsmith/index.html>):

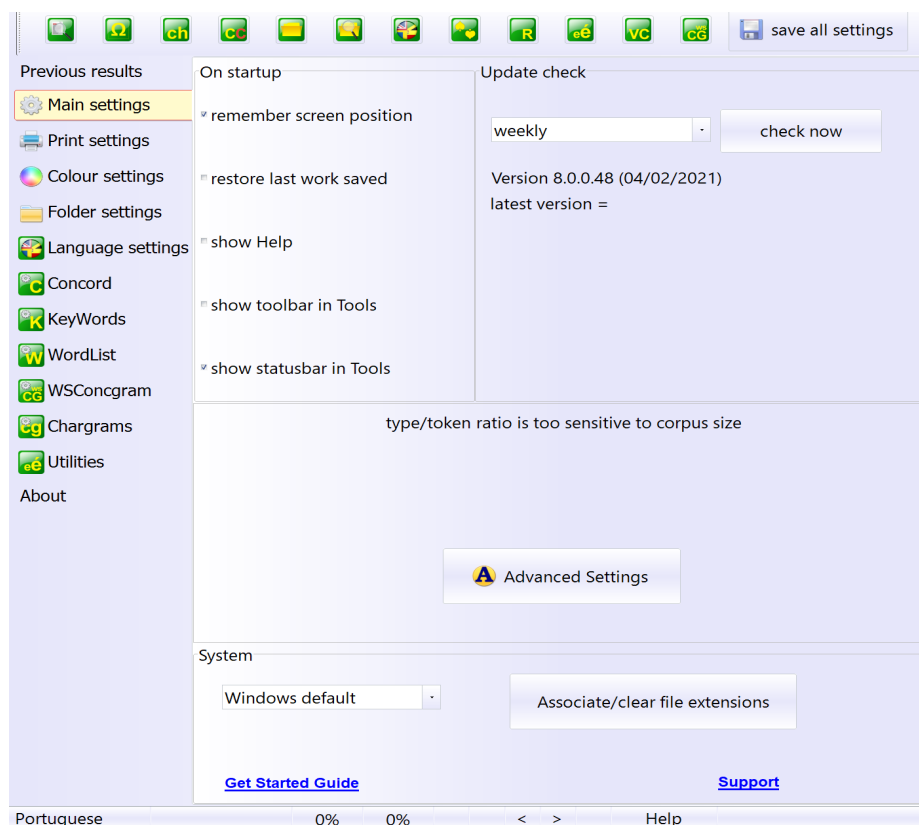


Figure 28. WordSmith Tool's main menu.



There are the three main tools, very common in other suites: wordlist, keywords and concordance. But WordSmith Tools has a lot of other smaller programs:

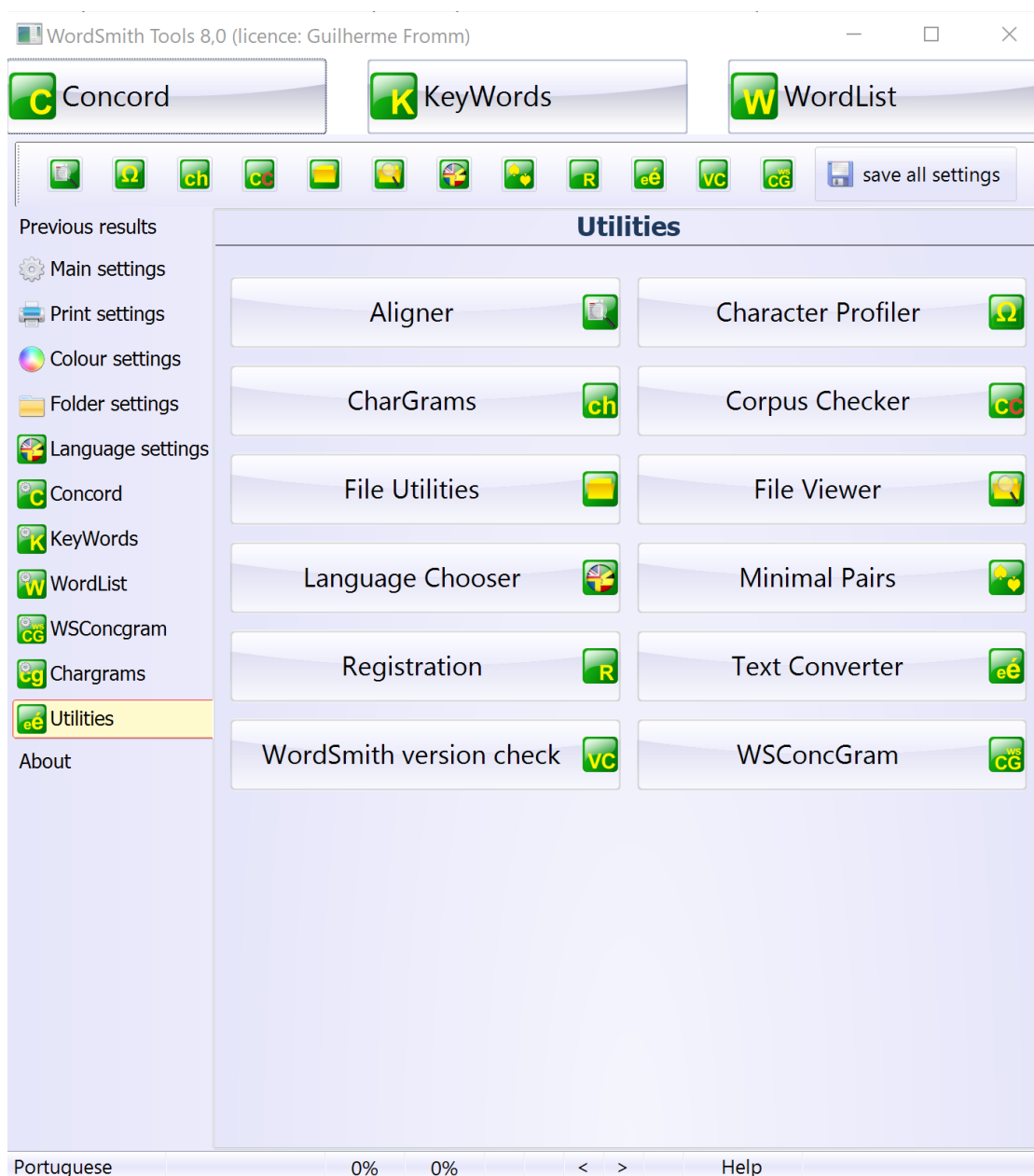


Figure 29. WordSmith Tools utilities.

In our course, we're going to work with a free option. The most famous one is the AntConc suite (<http://www.laurenceanthony.net/software.html>):

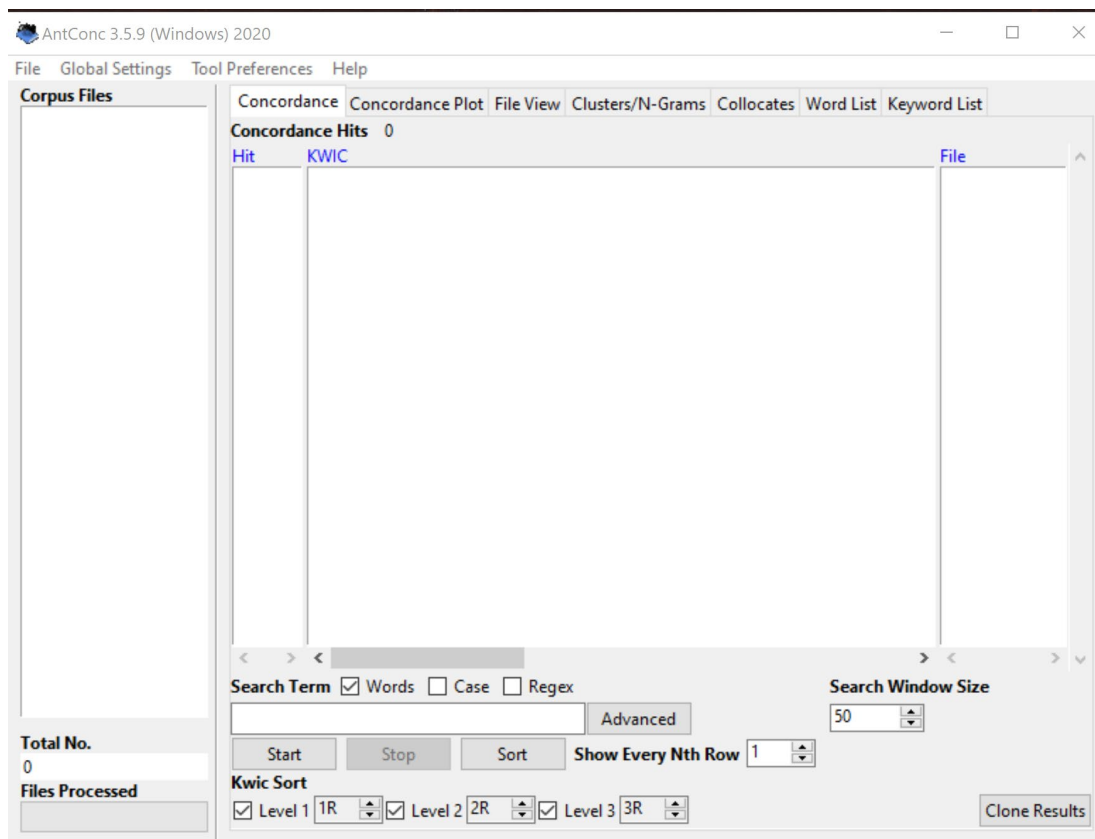


Figure 30. AntConc's main menu.

As AntConc is free, we're going to use this program to demonstrate the capabilities of describing language. But first, we must compile a corpus.

### Compiling a corpus

Following the idea presented on module 1, let's create our corpus. The summary of it:

Table 4. Corpus compilation.

<b>Language</b>	English
<b>Sources</b>	Academic Texts (articles, MA's and PhD's thesis)
<b>Time</b>	Synchronic
<b>Selection</b>	Study Corpus
<b>Content</b>	Specialized – Linguistics, Prosody
<b>Authorship</b>	Native and non-native speakers
<b>Internal Distribution</b>	monolingual
<b>Size</b>	Small
<b>Codification level</b>	Header

We propose you to create a corpus in a subfield of Linguistics, Prosody. Let's see the steps:

### Finding the texts

The best place to find texts is Google. But pay attention, a simple query wouldn't be enough. Use this kind of search:

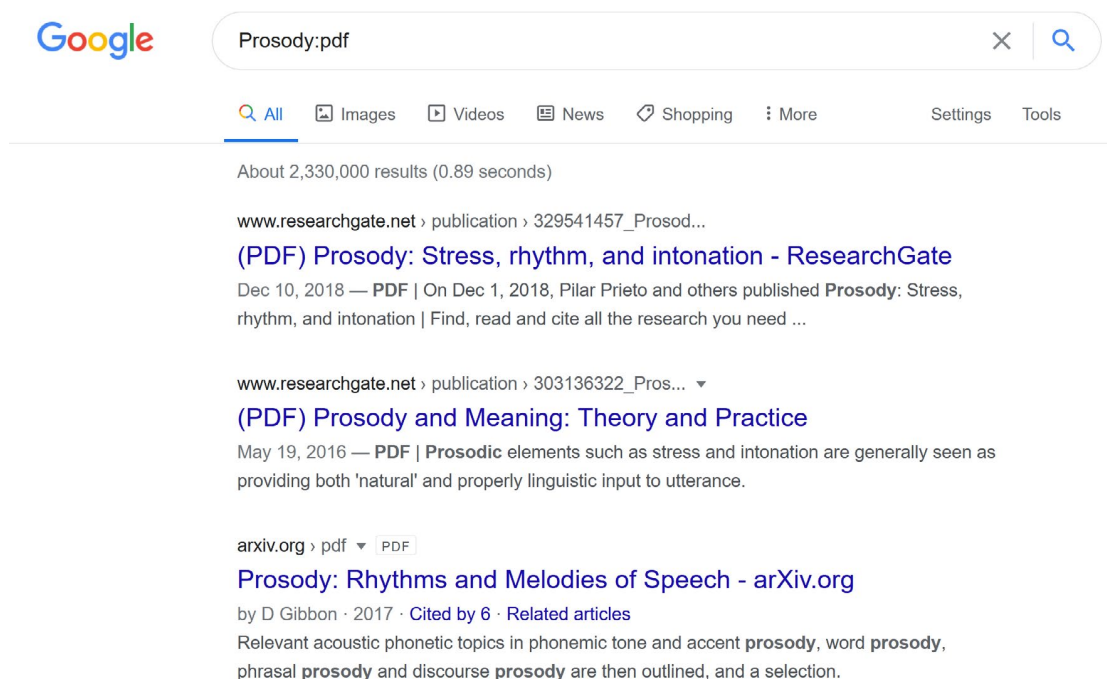
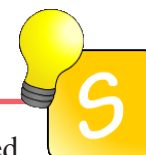


Figure 31. Google: PDF query for the word Prosody.

Take a look at the upper right corner of picture 22. Let's click the *options* menu and choose *advanced search*. We get the following screen:

We choose the pdf extension because the majority of articles (from specialized journals) and thesis available on the Internet are in this format.



Organization is very important when we work with Corpus Linguistics. To easily find the files you're going to work with, create a directory, in your computer, named Corpus Linguistics.



Let's take the third example. When you click the link, the site opens a file (named Prosody: Rhythms and Melodies of Speech). Before saving it, open your Corpus Linguistics directory and create a subdirectory named Corpus PDF. Save the file in it<sup>9</sup>. Notice that the file has 35 pages.

<sup>9</sup> Instead of using the file original name, try a shorter classification, like: File 1, File 2, etc.

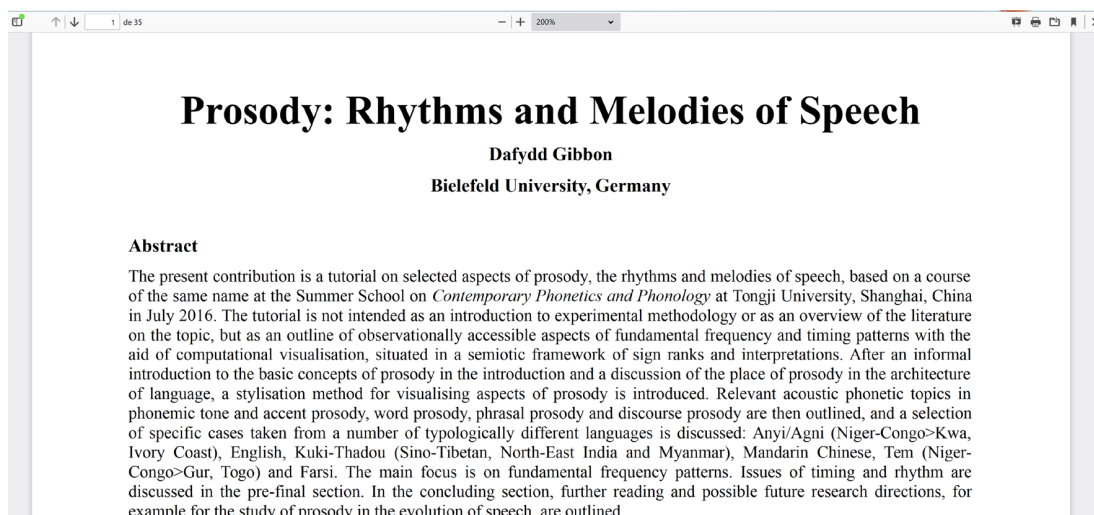


Figure 32. File Prosody: : *Rhythms and Melodies of Speech* first page.

In the sequence, save other files into your directory.

Something very important about the files you're going to use with lexical analysis tools is the format: the best way to work with the programs is to save your files in .txt format. Let's do the following: 1. Create a new subdirectory in your Corpus Linguistics directory, named Corpus TXT; 2. Open the Windows menu and choose All the programs > Windows Accessories > Notepad (figure 33); 3. open the Prosody: Rhythms and Melodies of Speech file, click the keys control + A together (to select the whole text, figure 34), copy it and paste it in the notepad file (figure 35).

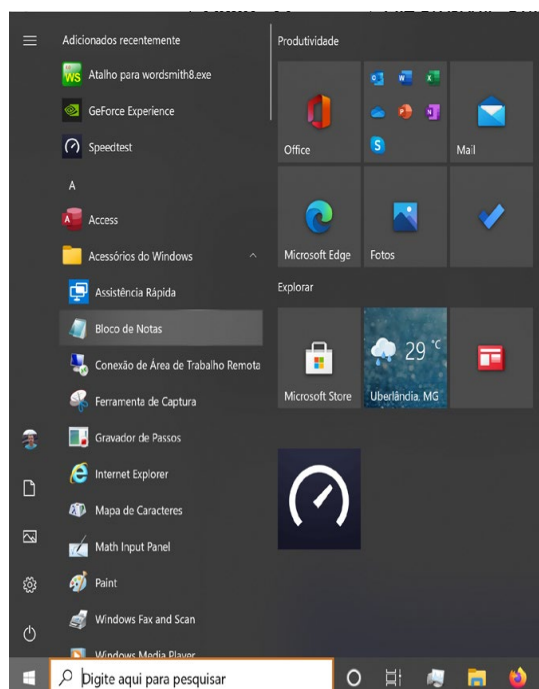


Figure 33. Notepad

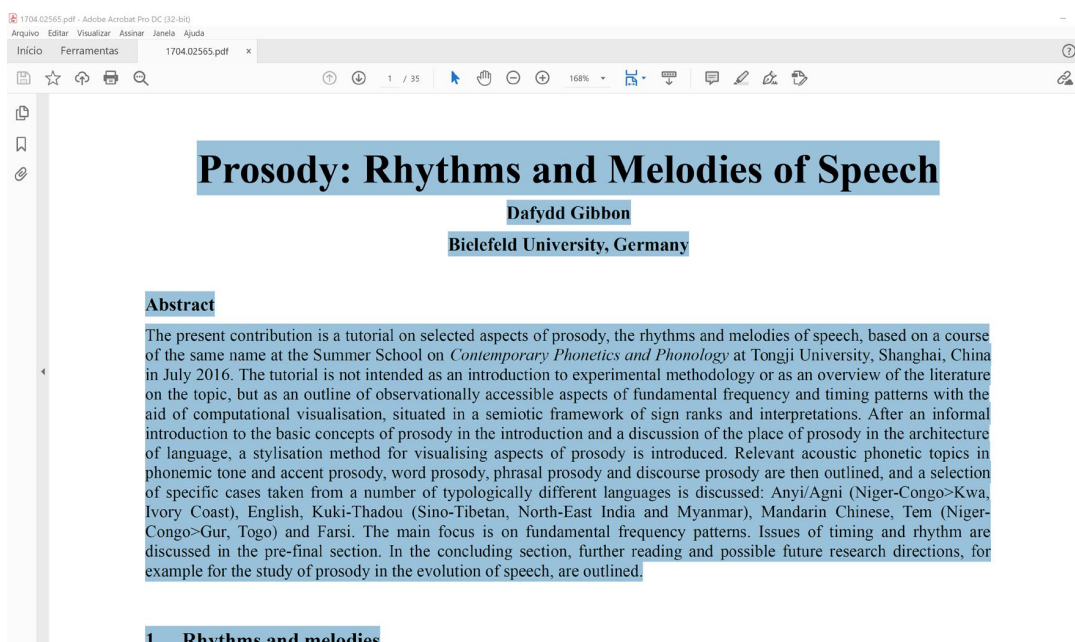


Figure 34. File *Prosody: Rhythms and Melodies of Speech* – whole text selected and ready to be copied.

As a result, you get this screen:

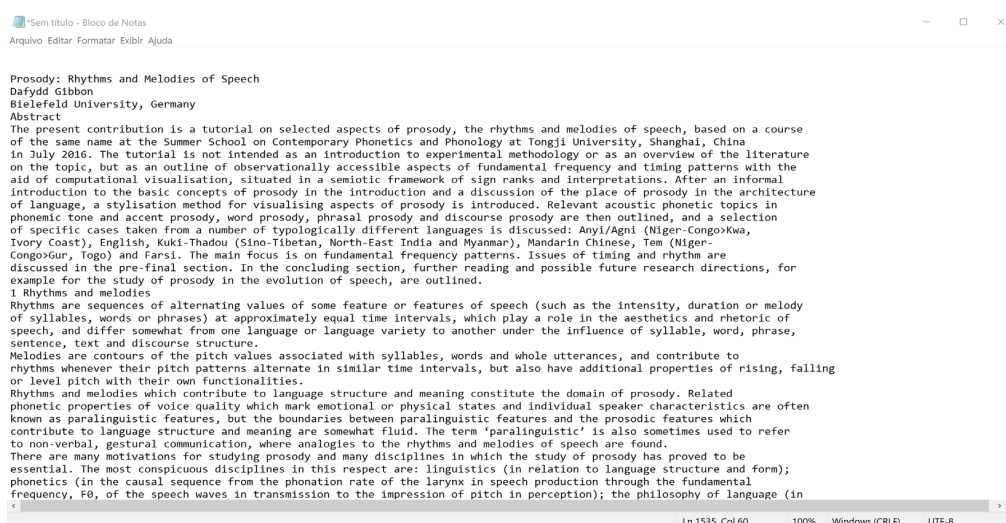


Figure 35. file *Prosody: Rhythms and Melodies of Speech* copied to the Notepad<sup>10</sup>.

Now we're going to insert a header (figure 36) for the file, containing its internet address and the date we collected the text. You're going to do the same procedure with all

The usage of angle brackets (< and >) is not for free. When the lexical analysis program reads the file, it ignores what is written inside these marks. We write the headers just for our information – they're not part of the text to be analyzed.



<sup>10</sup> As you can see, the Notepad erases all the formats, formulas, figures, etc. from the original text. Don't worry, because they're not important for corpus analysis.

Finally, you get the file complete:

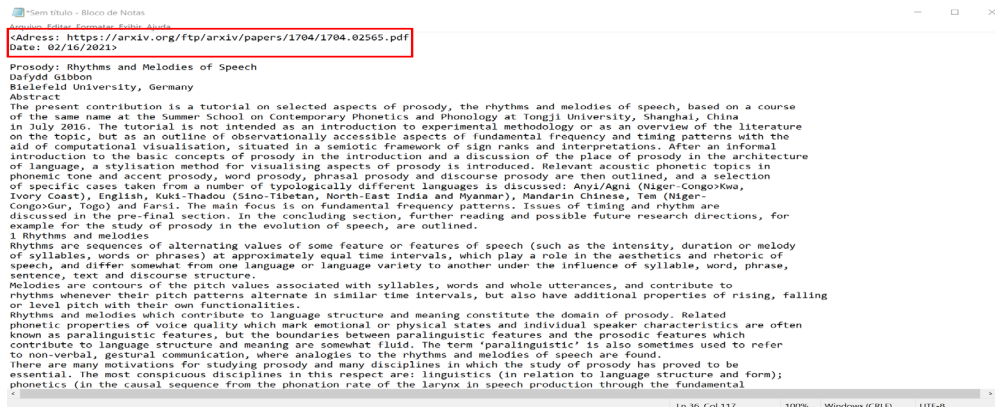


Figure 36. Text from file Prosody: *Rhythms and Melodies of Speech* copied to Notepad (with header).

The final step is saving the new file. Pay attention to some details: we're going to save the original file (Prosody: *Rhythms and Melodies of Speech*) as File 1:

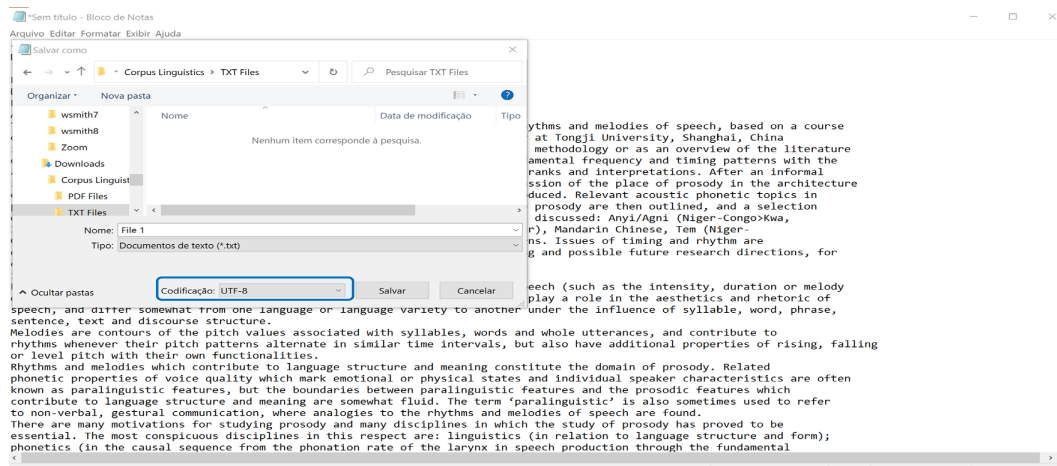




Figure 37. Saving the file Prosody: Rhythms and Melodies of Speech as File 1 and with UTF-8 codification.

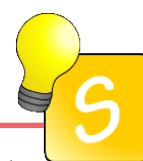
Pay attention to the code (codificação) we choose: UTF-8 (figure 37). It's very important to choose this one, because it works better with AntConc. If you want a more concise text, you can erase the pre (abstracts) and post (references) texts<sup>11</sup>.

<sup>11</sup> Generally, they're not important for your research.





Now you're going to finish compiling the corpus we've started above. The first text is already there (File 1). As soon as you finish compiling it, you are going to compact (into a .zip file) the directory (CORPUS TXT) and send it to your tutor. Your corpus must have, at least, 300,000 words<sup>12</sup>



Don't worry if you don't know how to count the words in different files. The lexical analysis tools will do it. We're going to learn how to use the AntConc in the sequence.

## AntConc

You've already seen the main menu of AntConc. As in all lexical analysis software, the first tool we're going to work with is the Word List.

### Word List

The first step to work with the Word List is to choose its tab and, in the sequence, open the directory with the files we're going to work with. In this case, let's open the Prosody directory (in your case, don't forget you saved your files in the Corpus TXT subdirectory):

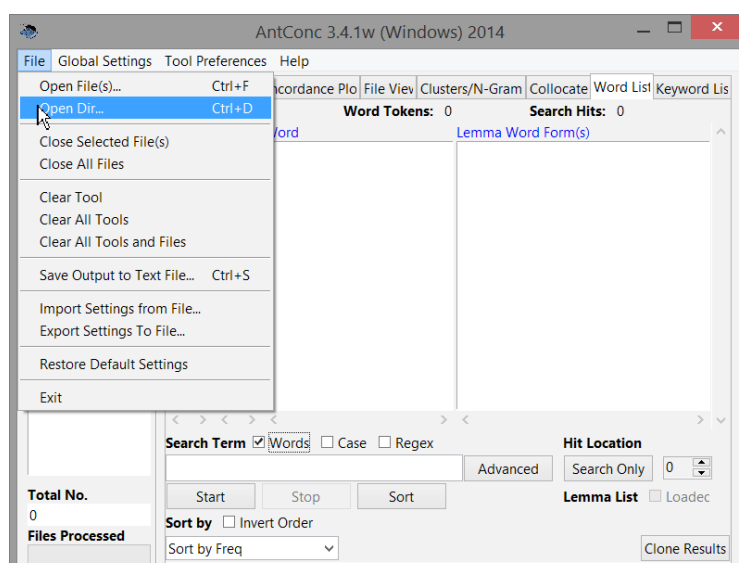


Figure 38. AntConc. Opening a directory in the WordList tool.

<sup>12</sup> You can use a text editor to calculate the size of your corpus. Put all of them in a single file and calculate it (in Microsoft Word, use Revision > Word Counting).

When you select to open the directory, you get the names of the files (included in this directory) you chose on the left field. In our case, the directory contains six files:

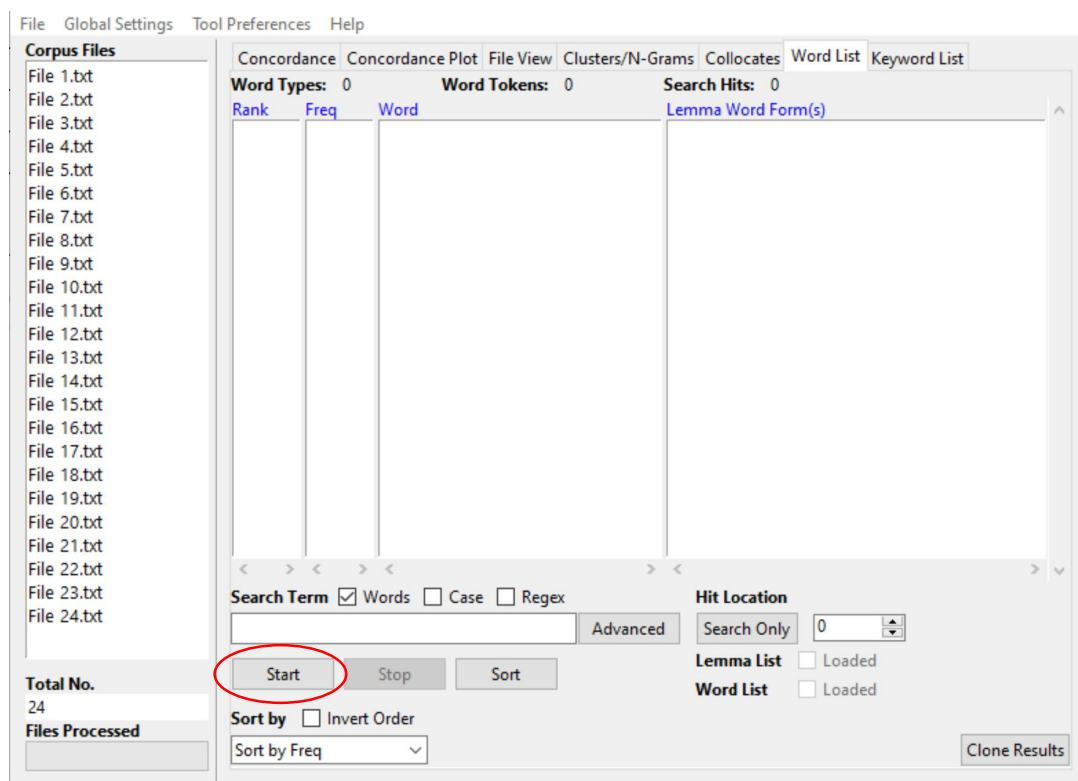


Figure 39. AntConc. Wordlist and files.

With the files opened, just click the START button (Figure 39). The program processes the corpus and gives you some information:

AntConc 3.5.9 (Windows) 2020

File Global Settings Tool Preferences Help

Corpus Files

Concordance Concordance Plot File View Clusters/N-Grams Collocates Word List Keyword List

Word Types: 17316 Word Tokens: 474727 Search Hits: 0

Rank	Freq	Word	Lemma	Word Form(s)
1	27095	the		
2	17162	of		
3	14652	a		
4	11127	and		
5	10759	to		
6	10676	in		
7	9288	is		
8	7428	that		
9	5694	e		
10	5122	s		
11	4773	t		
12	4329	i		
13	4297	as		
14	3841	for		
15	3560	be		
16	3324	it		
17	3321	are		
18	3112	this		
19	2992	n		
20	2954	not		
21	2778	or		
22	2733	with		
23	2580	o		
24	2579	by		
25	2449	r		
26	2400	on		
27	2340	an		
28	2120	which		

Total No. 24  
Files Processed

Search Term  Words  Case  Regex  Advanced

Hit Location Search Only 0

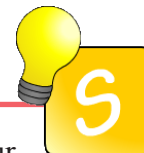
Lemma List  Loaded  
Word List  Loaded

Sort by  Invert Order  
Sort by Freq

Start Stop Sort Clone Results

Figure 40. AntConc. Corpus of Prosody, processed by the Word List tool.

In the middle of the screen, you find a lot of information: **Word types**, **Word Tokens**, and the **words** the corpus contains. What does all this information mean? It means we have a total corpus of 474,727 words (tokens), with 17,316 different words (types) and the most common word of this corpus is the article THE (it appears 27,095 times in the corpus).



These are the procedures you're going to follow when you are compiling your corpus and you want to know how many words you've already compiled.

Let's work a little bit more with the list. Suppose we're just looking for content words (nouns, adjectives and verbs). There's a way to reduce this list by using a Stoplist. A Stoplist contains words that you don't want the program to analyze. Usually we must prepare a list of words to achieve this goal, but here, we have already created one (download here: [https://drive.google.com/file/d/0B\\_pmpE08GOg2VIZVS0wzeTIyUzA/](https://drive.google.com/file/d/0B_pmpE08GOg2VIZVS0wzeTIyUzA/))



Don't forget to save this file in the Corpus Linguistics directory of your computer. Create a new subdirectory named Lists.

The first step in order to open a Stoplist is to access the Tool Preferences menu > Word List > Use a stoplist below > Open > Add words from file:

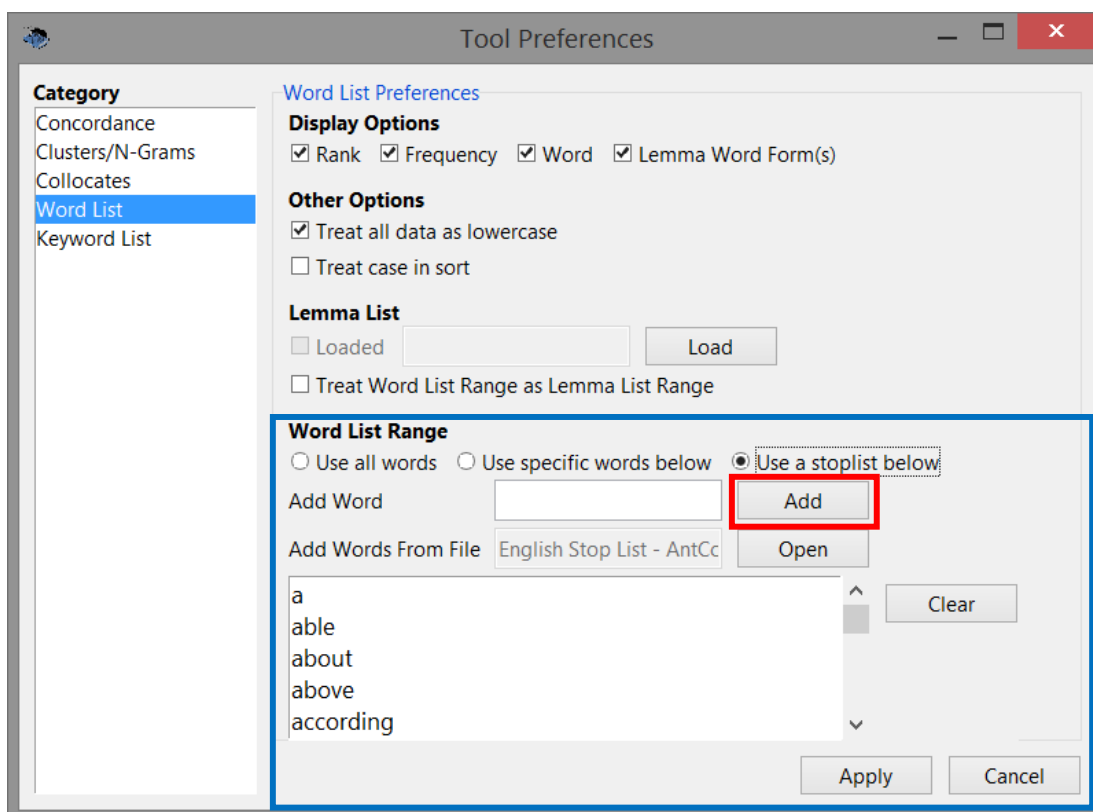


Figure 41. AntConc. Tool Preferences menu.

Get the file you've just downloaded (English Stop List – AntConc) and click APPLY. Now, run the Word List again (just press the START button). You get a new list:

AntConc 3.5.9 (Windows) 2020

File Global Settings Tool Preferences Help

Concordance Concordance Plot File View Clusters/N-Grams Collocates Word List Keyword List

Corpus Files

File 1.txt  
File 2.txt  
File 3.txt  
File 4.txt  
File 5.txt  
File 6.txt  
File 7.txt  
File 8.txt  
File 9.txt  
File 10.txt  
File 11.txt  
File 12.txt  
File 13.txt  
File 14.txt  
File 15.txt  
File 16.txt  
File 17.txt  
File 18.txt  
File 19.txt  
File 20.txt  
File 21.txt  
File 22.txt  
File 23.txt  
File 24.txt

Word Types: 16979 Word Tokens: 240640 Search Hits: 0

Rank	Freq	Word	Lemma Word Form(s)
1	9288	is	
2	3560	be	
3	3321	are	
4	1656	can	
5	1558	have	
6	1118	has	
7	1090	language	
8	1066	speaker	
9	1060	discourse	
10	1060	pragmatics	
11	1015	context	
12	971	meaning	
13	870	sentence	
14	865	speech	
15	849	was	
16	832	pragmatic	
17	827	will	
18	823	example	
19	815	use	
20	805	utterance	
21	782	may	
22	762	information	
23	731	like	
24	724	interpretation	
25	705	would	
26	676	do	
27	675	see	
28	642	been	

Search Term  Words  Case  Regex Hit Location Search Only 0

Start Stop Sort

Sort by  Invert Order Sort by Freq

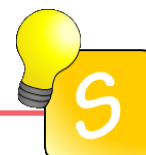
Lemma List  Loaded Word List  Loaded

Clone Results

Total No. 24  
Files Processed

Figure 42. Word List with Stoplist, Prosody.

As you can read above, the words, now, are more specific to the Prosody area. Notice that the tokens and types numbers have also changed. But there's still an empty field: Lemma Word Form.



Lemma is a basic form of a word. For example, the verb GO, in English, has the following forms: go, goes, going, went, gone. But when we're talking about the verb, we use its lemma, the infinitive GO. Get an English Lemma list here : [https://1drv.ms/u/s!Aq3R\\_KrvKKr\\_qrdDTwzWyiMz3Y2bKw?e=b45IsV](https://1drv.ms/u/s!Aq3R_KrvKKr_qrdDTwzWyiMz3Y2bKw?e=b45IsV)).

There's a way to associate all the forms of a word in the same line, using a Lemma List. Go back to the Tool Preferences menu > Wordlist > Lemma List. Load the list you've just downloaded (the list will be showed) and click APPLY.

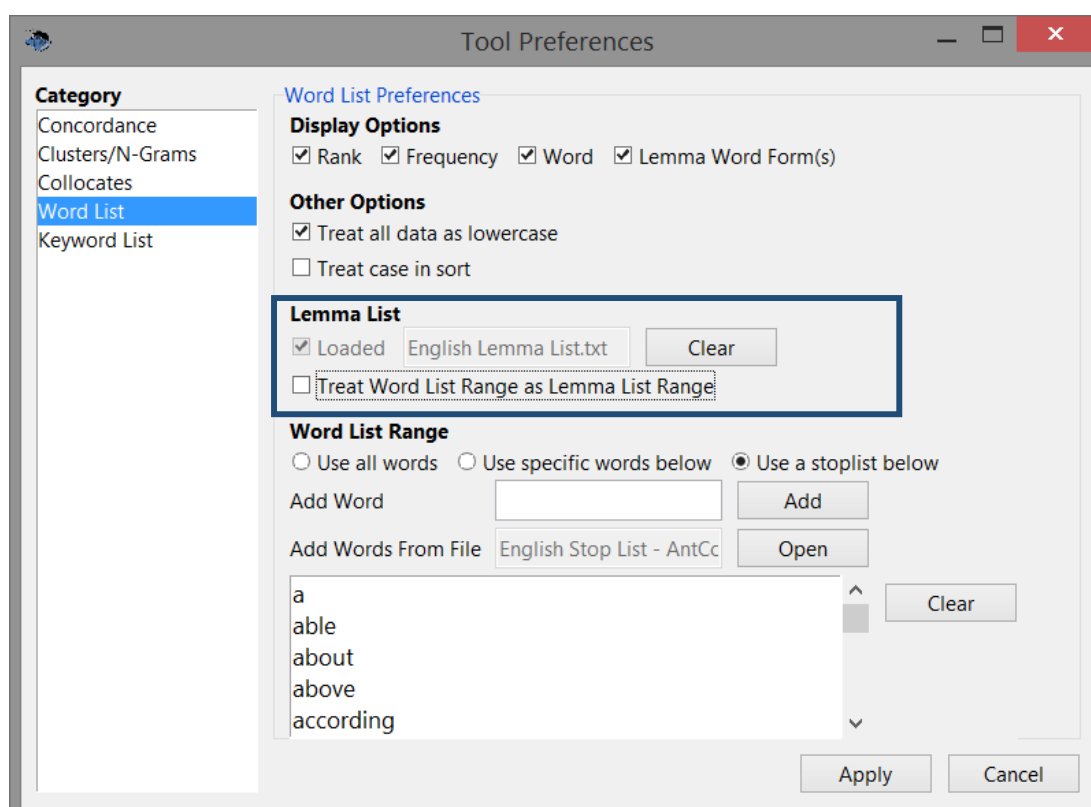


Figure 43. Loading the Lemma List.

You get a new list, with a Stop List and a Lemma List activated:



AntConc 3.5.9 (Windows) 2020

File Global Settings Tool Preferences Help

Corpus Files

File 1.txt  
File 2.txt  
File 3.txt  
File 4.txt  
File 5.txt  
File 6.txt  
File 7.txt  
File 8.txt  
File 9.txt  
File 10.txt  
File 11.txt  
File 12.txt  
File 13.txt  
File 14.txt  
File 15.txt  
File 16.txt  
File 17.txt  
File 18.txt  
File 19.txt  
File 20.txt  
File 21.txt  
File 22.txt  
File 23.txt  
File 24.txt

Concordance Concordance Plot File View Clusters/N-Grams Collocates Word List Keyword List

Lemma Types: 13298 Lemma Tokens: 240640 Search Hits: 0

Rank	Freq	Lemma	Lemma Word Form(s)
1	18033	be	are 3321 be 3560 been 642 is 9288 was 849 were 373
2	3084	have	had 305 has 1118 have 1558 having 103
3	1892	pragmatic	pragmatic 832 pragmatics 1060
4	1779	use	use 815 used 588 uses 176 using 200
5	1656	can	can 1656
6	1365	speaker	speaker 1066 speakers 299
7	1357	language	language 1090 languages 267
8	1351	sentence	sentence 870 sentences 478 sentencing 3
9	1290	example	example 823 examples 467
10	1260	context	context 1015 contexts 245
11	1196	meaning	meaning 971 meanings 225
12	1088	do	did 271 do 676 doing 93 done 48
13	1087	discourse	discourse 1060 discourses 27
14	1076	utterance	utterance 805 utterances 271
15	977	semantic	semantic 505 semantics 472
16	927	say	said 408 say 378 says 141
17	905	act	act 482 acted 2 acting 10 acts 411
18	902	see	saw 39 see 675 seeing 26 seen 149 sees 13
19	870	speech	speech 865 speeches 5
20	827	will	will 827
21	818	case	case 482 cases 336
22	816	interpretation	interpretation 724 interpretations 92
23	782	may	may 782
24	762	information	information 762
25	759	give	gave 51 give 174 given 447 gives 51 giving 36
26	746	like	like 731 liked 3 likes 12
27	730	linguistic	linguistic 562 linguistics 168
28	718	expression	expression 313 expressions 405

Search Term  Words  Case  Regex

Hit Location Search Only 0

Lemma List  Loaded

Word List  Loaded

Total No. 24

Files Processed

Sort by  Invert Order

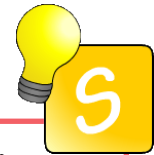
Sort by Freq

Clone Results

Figure 44. Word List with Stoplist and Lemma list.

The usage of Lemma lists is not always interesting for your study. For example, in figure 44, the word ACT (905) contains: act (482), acted (2), acting (10) and acts (411). However, there's a catch: how do you know if act or acts are nouns or verbs<sup>13</sup>? This means that, as already explained before, you must plan very well the kind of research you're doing.

<sup>13</sup> The concordance lines can tell you the difference. But you need to analyse one by one.



There's a big disadvantage of using AntConc (remember, is a free suite): you can't save your results. The only possibility, in the menu, is to click File > Save Output to Text File.

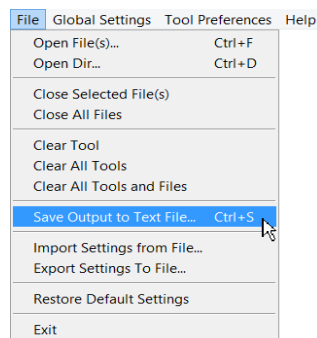


Figure 45. Saving file in AntConc.

As a result, you get a file like this one:

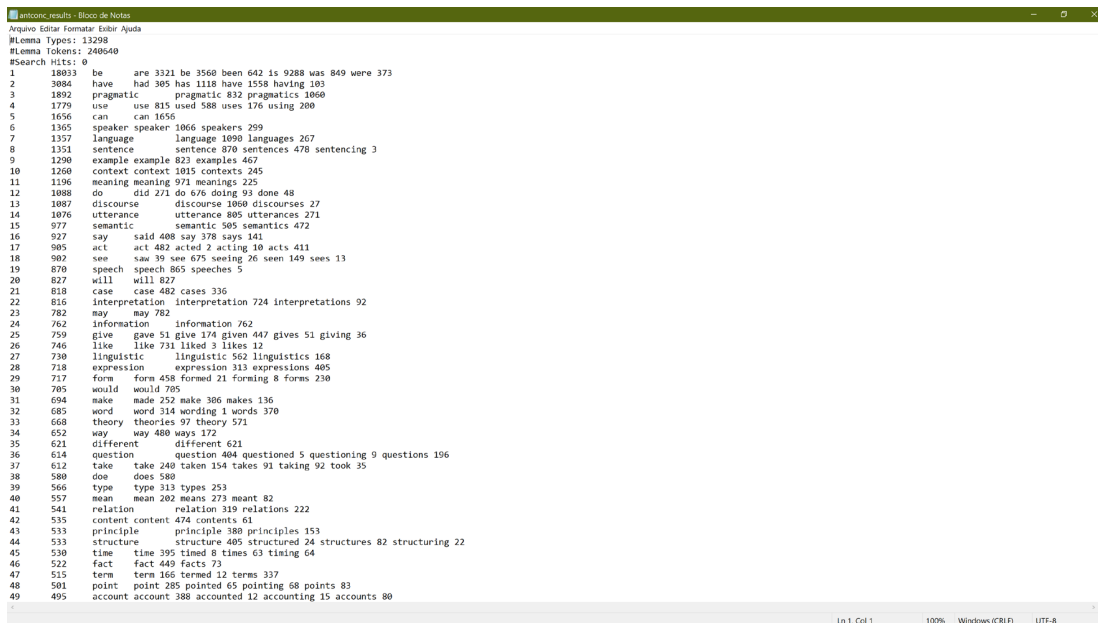


Figure 46. File saved from AntConc's Word List.

## Concordance

To obtain the concordance lines of a word, you just need to click it. For example, click the word PRAGMATIC (see figure 44). As a result, you get this screen:

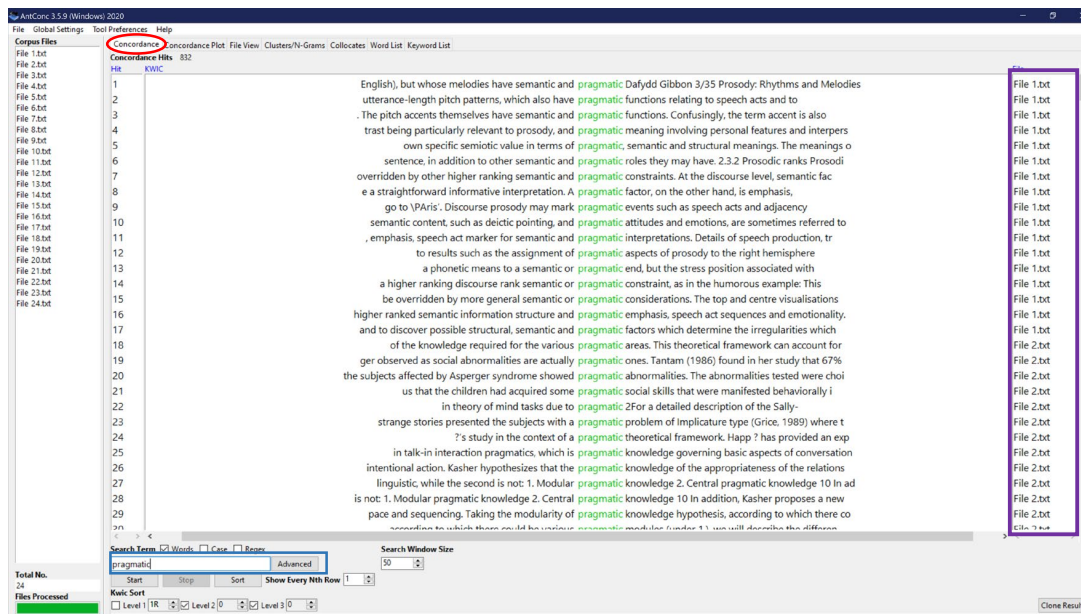


Figure 47. Concordances lines for the word PRAGMATIC.

As you can see, the program displays a list of every entry of the word PRAGMATIC (centered) in a KWIC screen (and the word appears selected as a search). On the right side, you see the name of the file the example it comes from. If you click on the result in green, you get the whole context (and the word you chose is highlighted) in the File View tab. Let's try it with the first line:

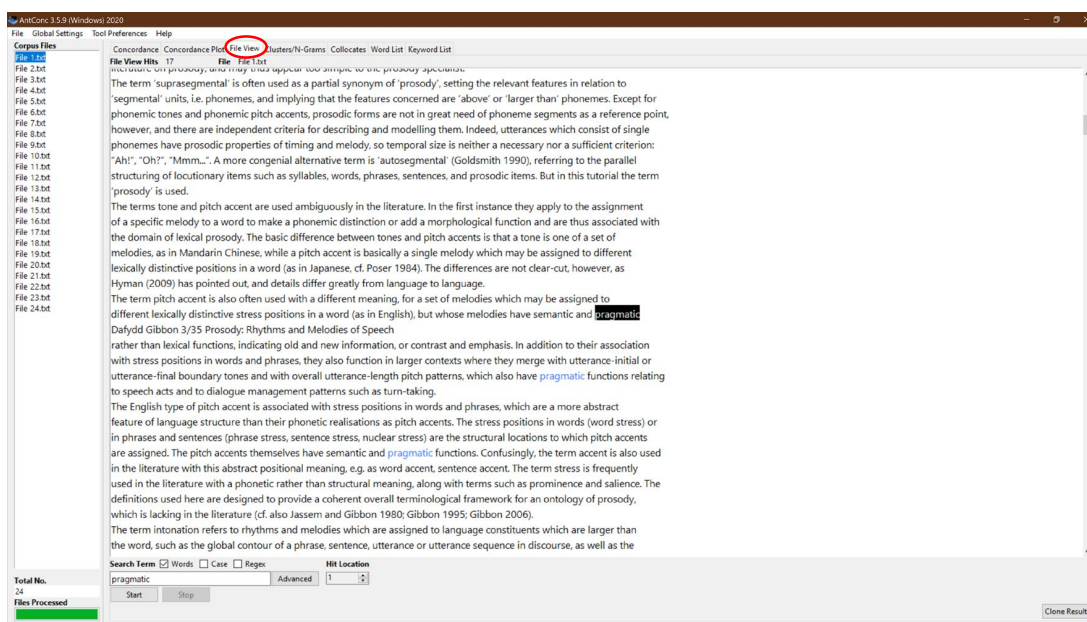


Figure 48. File View for the first entry of the word PRAGMATIC.

If you want, you can copy the text using the control + c keys. Another possibility of viewing the behavior of the word PRAGMATIC is by clicking the tab Concordance Plot:

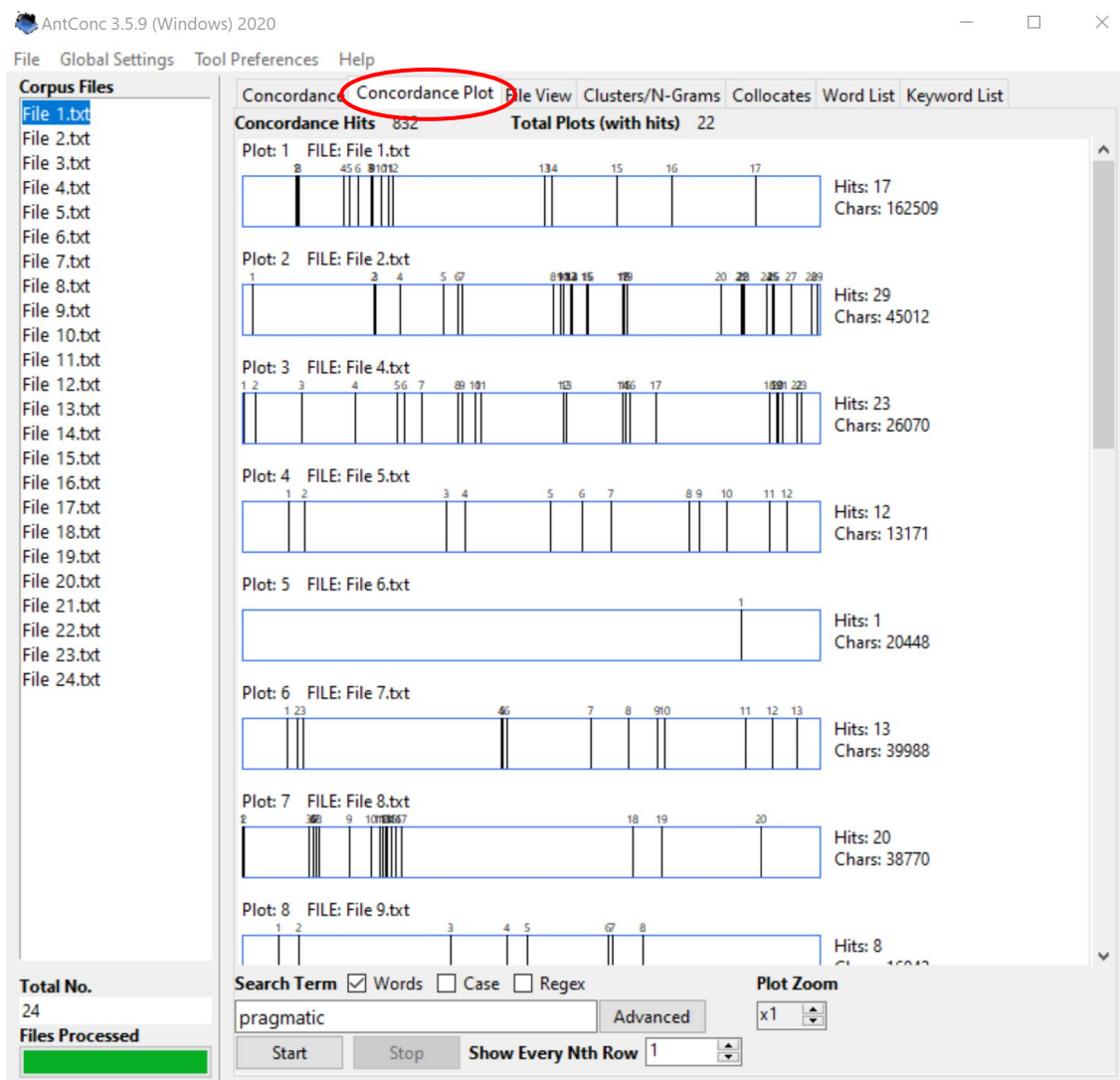


Figure 49. Concordance Plot for the word PRAGMATIC.

The slashes you see in the rectangles represent the distribution of the word PRAGMATIC in each text of the corpus.

PRAGMATIC, as you noticed, is an adjective. The name of the area is PRAGMATICS (a noun). Let's search it? You just need to go back to the Concordance tab and write the word in the [search field](#) (as in figure 47) and click START. Then we get:



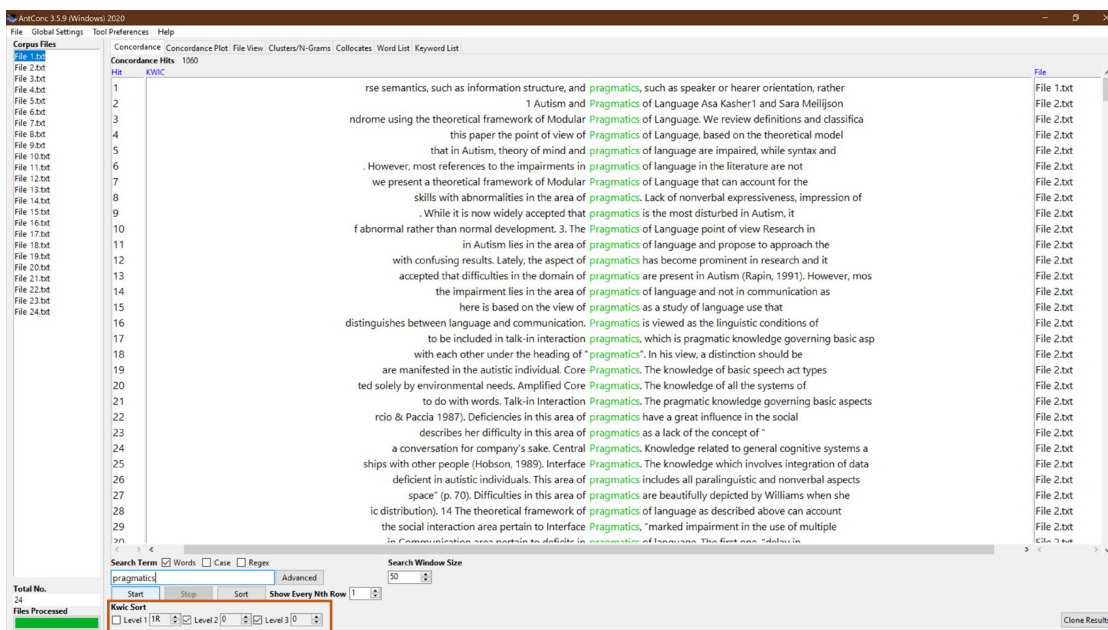


Figure 50. Choosing the word PRAGMATICS for concordance lines.

As you can see in figure 50, there are **parameters** you can change to show different results. Let's try with this configuration: **Kwic Sort**  Level 1 **1R**  Level 2 **1L**  Level 3 **0** (remember: choose your configurations and click the START button). The result we get is this:

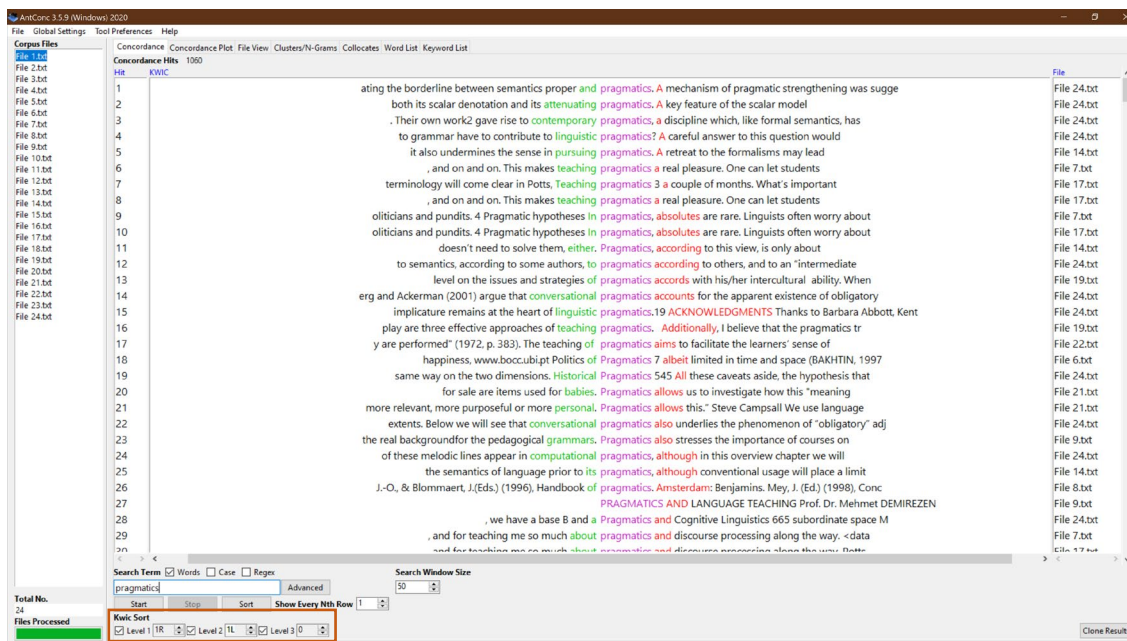
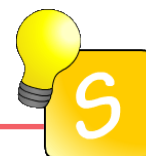


Figure 51. Different levels of sorting with PRAGMATICS concordance lines.



The parameters we choose must reflect our research. In this case, we're just playing with the tool, but for a specific research you need to choose specific parameters.

Yet another possibility of analyzing the word you are interested in is to search the words that co-occur with it. Go back to Tools Preferences and erase the configurations we did for the StopList and LemmaList, and then load the WordList again (with no search term); after that, write the word SPEAKER in the same field and click the button search only (figure 52).

The screenshot shows the AntConc 3.5.9 (Windows) 2020 interface. The 'Word List' tab is active, displaying a list of words and their frequencies. The search term 'speaker' is entered in the 'Search Term' field, and the 'Search Hits' are 1. The 'Hit Location' is set to 'Search Only'.

Rank	Freq	Word	Lemma	Word Form(s)
36	1582	what		
37	1558	have		
38	1410	one		
39	1373	I		
40	1371	they		
41	1306	at		
42	1296	f		
43	1285	m		
44	1251	there		
45	1248	d		
46	1164	p		
47	1118	has		
48	1090	language		
49	1066	speaker		
50	1060	discourse		
51	1060	pragmatics		
52	1043	these		
53	1032	other		
54	1015	context		
55	981	such		
56	980	some		
57	972	if		
58	971	meaning		
59	955	g		
60	938	he		
61	928	more		

Search Term:  Words  Case  Regex  
 speaker  
 Hit Location: Search Only 1  
 Lemma List  Loaded  
 Word List  Loaded  
 Sort by:  Invert Order  
 Sort by Freq

Figure 52. Searching for a specific word in the WordList.

To continue with the analysis of the word SPEAKER, now we're going to work with the tab Collocates. Just choose the tab, certify that the word SPEAKER is still in the search field and click Start. The result you can see in figure 53 (in this case, we used the **sort by frequency** option).

AntConc 3.5.9 (Windows) 2020

File Global Settings Tool Preferences Help

Corpus Files

File 1.txt  
File 2.txt  
File 3.txt  
File 4.txt  
File 5.txt  
File 6.txt  
File 7.txt  
File 8.txt  
File 9.txt  
File 10.txt  
File 11.txt  
File 12.txt  
File 13.txt  
File 14.txt  
File 15.txt  
File 16.txt  
File 17.txt  
File 18.txt  
File 19.txt  
File 20.txt  
File 21.txt  
File 22.txt  
File 23.txt  
File 24.txt

Concordance Concordance Plot File View Clusters/N-Grams **Collocates** Word List Keyword List

Total No. of Collocate Types: 1875 Total No. of Collocate Tokens: 10660

Rank	Freq	Freq(L)	Freq(R)	Stat	Collocate
1	1230	923	307	4.33745	the
2	377	168	209	3.96391	to
3	349	31	318	4.92334	s
4	348	230	118	3.17476	of
5	343	214	129	3.38200	a
6	266	182	84	3.99527	that
7	260	111	149	3.63996	is
8	248	83	165	3.31117	and
9	196	94	102	3.03137	in
10	101	74	27	4.82942	what
11	100	70	30	4.11001	by
12	98	38	60	3.97363	or
13	94	27	67	6.89801	hearer
14	79	30	49	3.40382	it
15	75	29	46	3.49911	not
16	74	38	36	3.59193	with
17	74	17	57	7.29395	addressee
18	73	29	44	2.91946	as
19	72	40	32	3.06141	for
20	69	22	47	3.10961	be
21	61	38	23	3.50067	on
22	60	12	48	4.78231	meaning
23	58	36	22	4.94937	about
24	52	14	38	4.37248	has
25	50	21	29	2.83898	this
26	47	29	18	3.30349	which
27	45	20	25	3.00024	...

Search Term  Words  Case  Regex  
speaker Advanced

Window Span  Same  
From... 5L To... 5R

Min. Collocate Frequency  
1

Start Stop Sort

Sort by  Invert Order  
Sort by Freq

Total No. 24  
Files Processed

Clone Results

Figure 53. Collocates for the word SPEAKER.

Observe the example of the word HEARER: it occurs 94 times in the corpus, 27 times to the left of HEARER and 67 times to the right of it. If you click the word **HEARER** in this screen, you go back to the Concordance tab with these combinations and you get



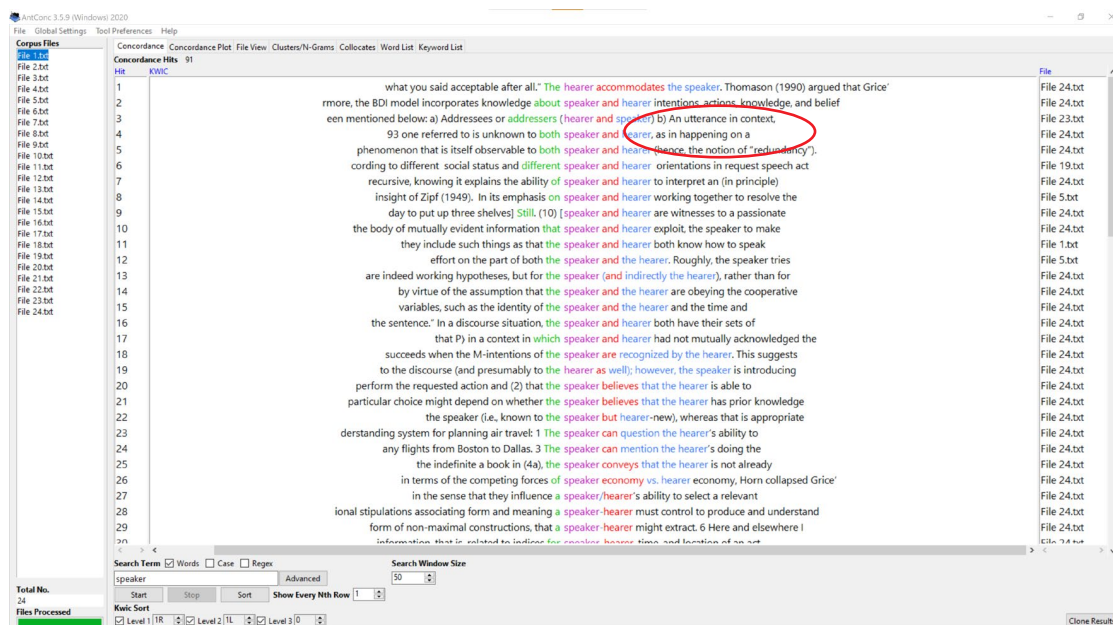


Figure 54. Collocates for SPEAKER + HEARER.

The final tab related to the Concordance tool is the Clusters/N-Grams. Still using the same screen, click on the tab CLUSTERS/N-Grams and the START button:

AntConc 3.5.9 (Windows) 2020

File Global Settings Tool Preferences Help

Corpus Files

File 1.txt  
File 2.txt  
File 3.txt  
File 4.txt  
File 5.txt  
File 6.txt  
File 7.txt  
File 8.txt  
File 9.txt  
File 10.txt  
File 11.txt  
File 12.txt  
File 13.txt  
File 14.txt  
File 15.txt  
File 16.txt  
File 17.txt  
File 18.txt  
File 19.txt  
File 20.txt  
File 21.txt  
File 22.txt  
File 23.txt  
File 24.txt

Concordance Concordance Plot File View **Clusters/N-Grams** Collocates Word List Keyword List

Total No. of Cluster Types 2989 Total No. of Cluster Tokens 4264

Rank	Freq	Range	Cluster
1	255	12	speaker's
2	56	6	speaker is
3	49	7	speaker and
4	32	7	speaker to
5	30	6	speaker has
6	25	2	speaker's meaning
7	23	1	speaker means
8	21	2	speaker's intention
9	18	4	speaker of
10	16	1	speaker meaning
11	15	1	speaker and addressee
12	14	2	speaker who
13	14	1	speaker's empathy
14	13	2	speaker intends
15	13	1	speaker says
16	12	3	speaker may
17	12	5	speaker or
18	11	4	speaker and hearer
19	11	2	speaker intends to
20	11	3	speaker must
21	11	1	speaker's empathy with
22	11	1	speaker's intentions
23	11	4	speaker's utterance
24	10	1	speaker does
25	10	3	speaker intentions
26	10	2	speaker wants
27	10	2	speaker's communicative
28	10	3	speaker's intended

Search Term  Words  Case  Regex  N-Grams

speaker Advanced

Start Stop Sort

Cluster Size Min. 2 Max. 5

Min. Freq. Min. Range 1 1

Sort by  Invert Order Search Term Position  On Left  On Right

Sort by Freq

Total No. 24 Files Processed

Clone Results

Figure 55. Clusters for the word SPEAKER.

In this case, we have a **cluster size** with 2 to 5 (from 2 to 5 words). We can change it, for example, to 2 to 4:

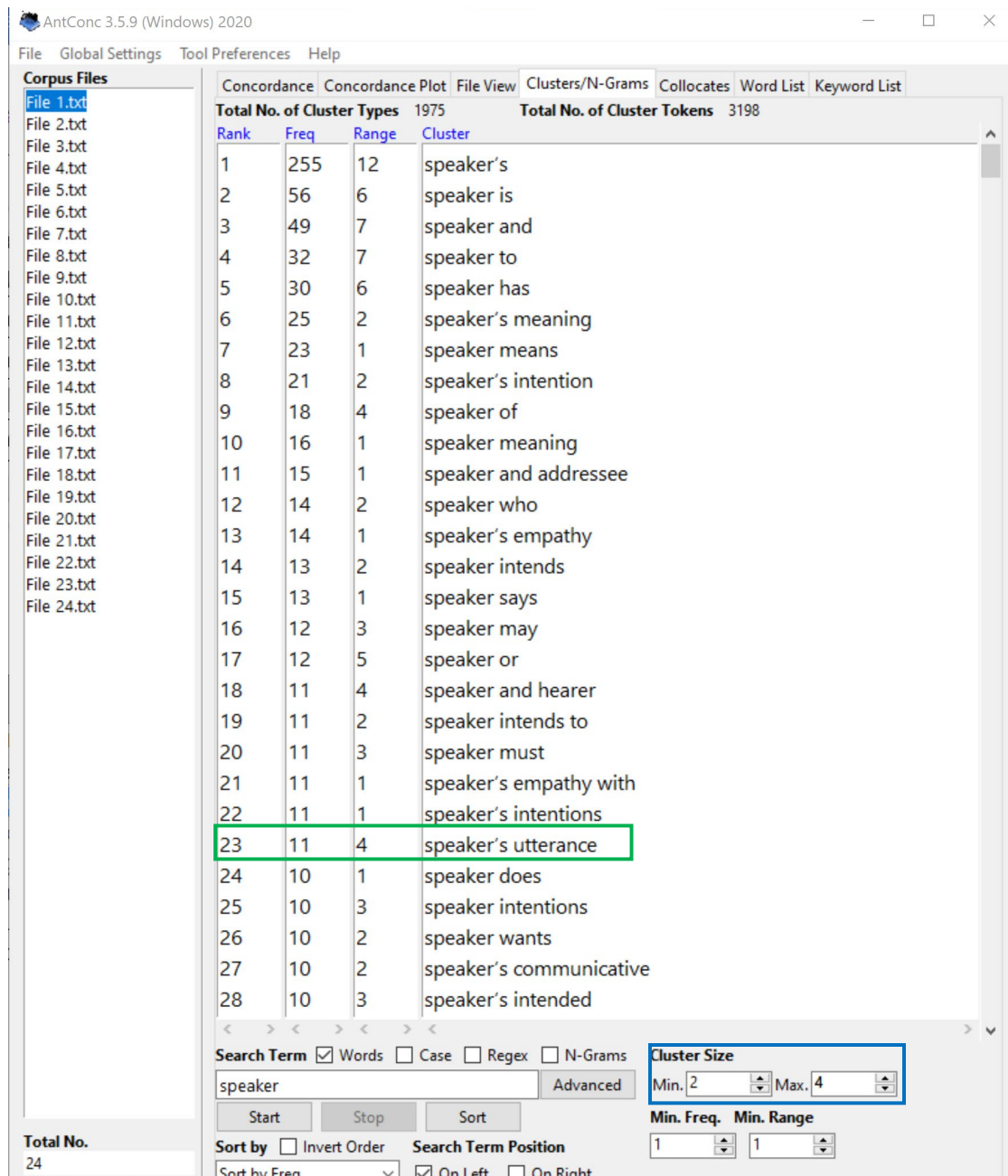


Figure 56. Clusters for the word SPEAKER.

We can search the cluster **SPEAKER'S UTTERANCE**, for example (note that we changed the Search Window Size parameters to 100, to visualize larger fragments of text):

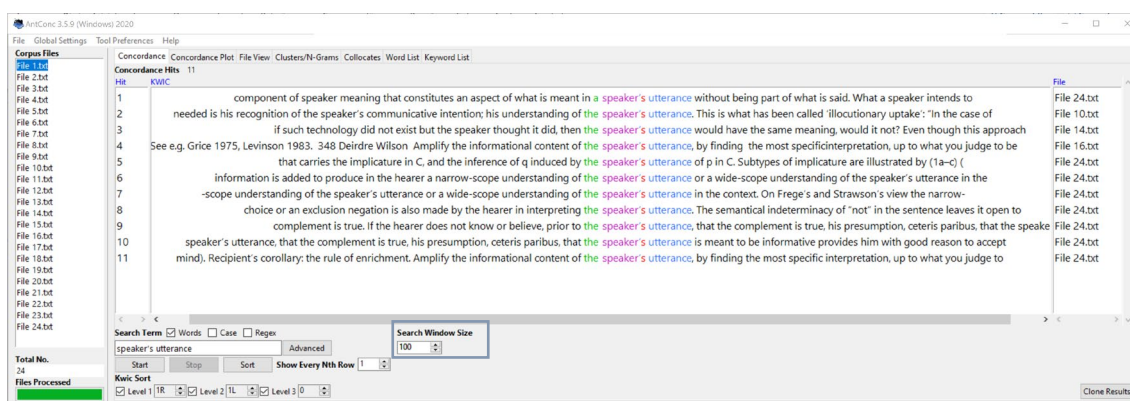


Figure 57. Concordance lines for the cluster SPEAKER'S UTTERANCE.

Remember, the lexical analysis programs exist to help, not to do the entire work for you. In figure 57, for example, you can start analyzing the SPEAKER'S UTTERANCE cluster for your research.



### Keywords

To find the keywords of a corpus, you need to have two available corpora: your study corpus and a reference corpus. You can download a reference corpus list from here: [https://1drv.ms/u/s!Aq3R\\_KrvKKr\\_qrdDTwzWyiMz3Y2bKw?e=b45IsV](https://1drv.ms/u/s!Aq3R_KrvKKr_qrdDTwzWyiMz3Y2bKw?e=b45IsV)

The first step is to load your reference corpus. Go to the upper menu of AntConc again and choose Tool Preferences > Keyword List > Reference Corpus > Use Word List > Add File. Get the file you've just downloaded and click load. The name of the corpus will appear in the rectangle. After that, click Load. The program will load the reference corpus and then will show a green rectangle, showing that everything is ok. Click APPLY to go

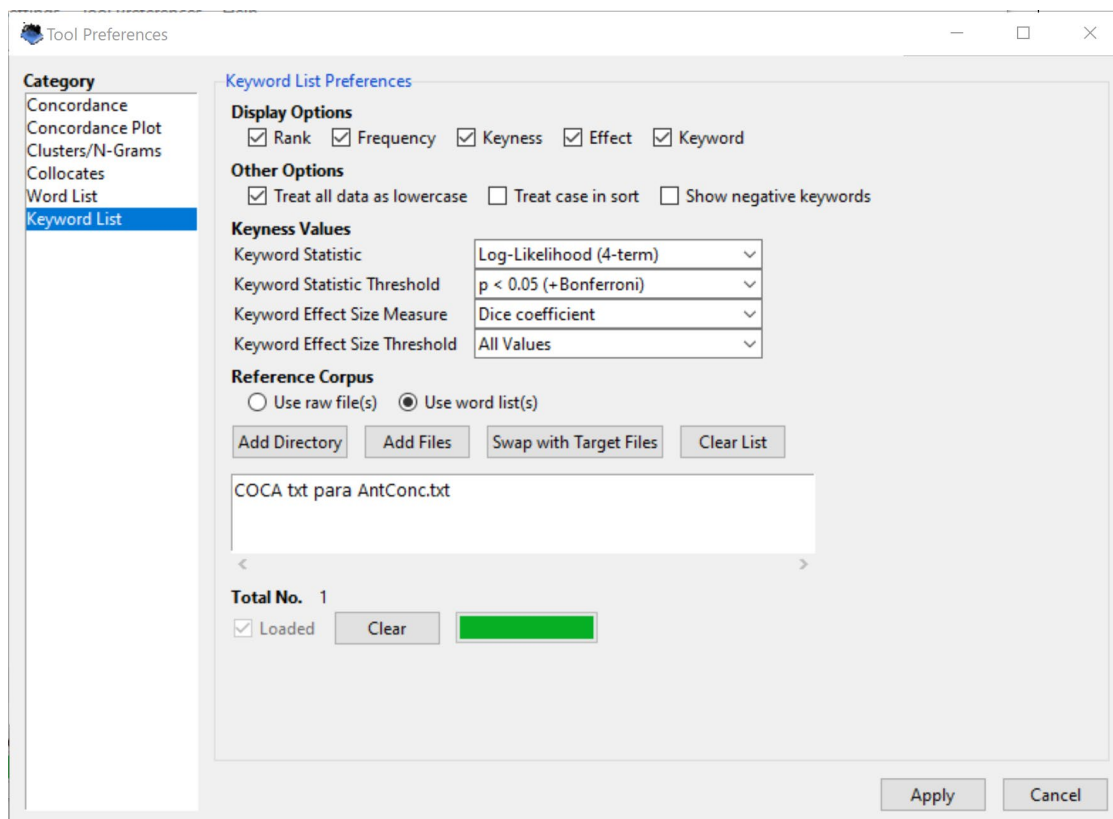


Figure 58. AntConc's Tool Preferences and Keyword List.

You have already changed the settings and loaded the reference corpus. Just choose the Keyword List tab of AntConc and click the START button. You get your keywords list from the PRAGMATICS area:

AntConc 3.5.9 (Windows) 2020

File Global Settings Tool Preferences Help

Concordance Concordance Plot File View Clusters/N-Grams Collocates Word List **Keyword List**

Keyword Types: 5687 Keyword Tokens: 254006 Search Hits: 0

Rank	Freq	Keyness	Effect	Keyword
1	1060	+ 13792.5	0.0088	pragmatics
2	1656	+ 13327.94	0.0132	can
3	870	+ 11074.27	0.0072	sentence
4	805	+ 8965.74	0.0066	utterance
5	971	+ 7978.52	0.0079	meaning
6	827	+ 7908.47	0.0068	will
7	1060	+ 7522.3	0.0084	discourse
8	731	+ 7447.96	0.006	like
9	832	+ 7431.83	0.0068	pragmatic
10	1066	+ 6925.03	0.0083	speaker
11	478	+ 6267.82	0.004	sentences
12	421	+ 5768.35	0.0035	john
13	425	+ 5456.1	0.0035	new
14	870	+ 5336.89	0.0068	sentence
15	482	+ 5208.62	0.004	case
16	1015	+ 5112.34	0.0075	context
17	474	+ 5053.77	0.0039	content
18	472	+ 4954.71	0.0039	semantics
19	971	+ 4895.23	0.0072	meaning
20	349	+ 4781.73	0.0029	grice
21	370	+ 4698.95	0.0031	words
22	405	+ 4674.79	0.0034	structure
23	364	+ 4462.35	0.003	even
24	505	+ 4445.39	0.0042	semantic
25	420	+ 4358.09	0.0035	reference
26	562	+ 4314.39	0.0046	linguistic
27	724	+ 4279.76	0.0057	interpretation

Search Term  Words  Case  Regex  Advanced Hit Location Search Only 0

Start Stop Sort Reference Corpus  Loaded

Sort by  Invert Order Sort by Keyness

Total No. 24 Files Processed

Clone Results

Figure 59. Keyword List from the Prosody corpus.

From this point on, you already know the procedures: click the word you want to analyze and you get its concordance lines. For example, let's click the word DISCOURSE (rank 7) and see the results:



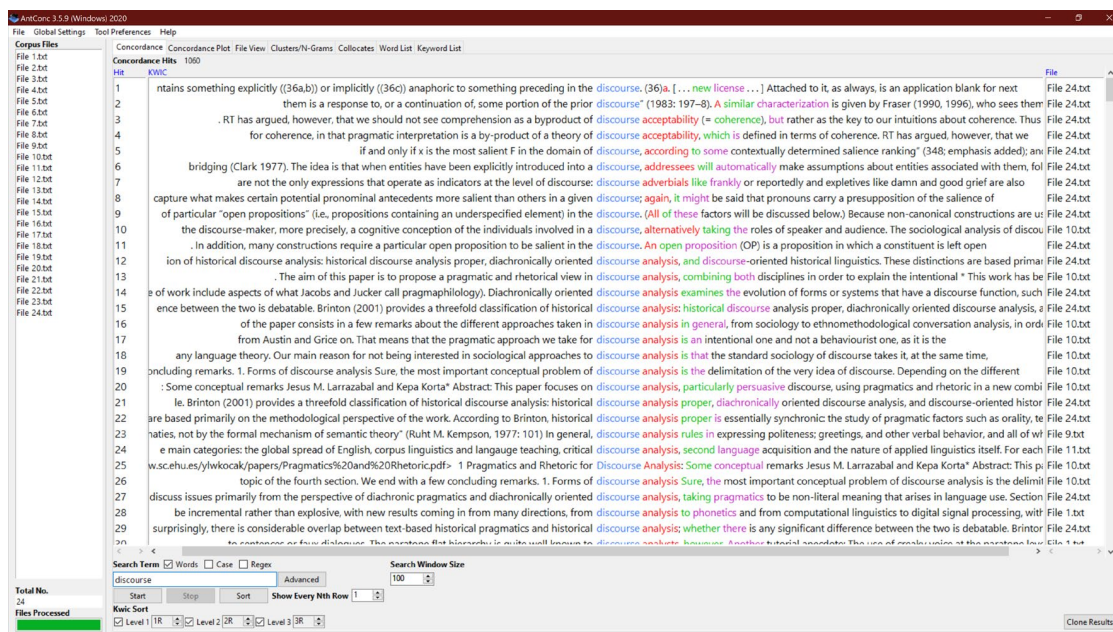


Figure 60. Concordance lines for the word DISCOURSE

From here, you can try the tabs we have already explained, like Concordance Plot, Collocates and Clusters/N-Grams. Have fun!



# MÓDULO 4

## Project Development

### **Basic Contents**

- Describing language nowadays.
- Choosing an area to be described: Terminography.
- Web environment for terminological management: VoTec.

### **Objectives**

- Recognize the steps of a language description project.
- Prepare a project you can use with your students.

## PROJECT DEVELOPMENT

---

### ACTIVITY 11



Video class, module 4. Watch the professor's hints about the subjects that are going to be worked in this module.

### ACTIVITY 12



Project. You're going to prepare a class plan project, using the Corpus Linguistics approach. Possible sources: Internet sites with corpora and tools (module 2), corpus compilation (module 3), lexical analysis tools software (module 3), a combination of some or all this previous ideas. Here in module 4 we present a possible project you could work with your students. Think about a project you really could apply in your school.

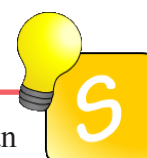
The project must contain all the steps you intend to develop with the students (don't forget to specify the target public, the description of the school you plan to apply the project, the available resources in this school, etc.). You've seen in this guide that pictures help a lot to prepare this kind of work – use them. When you finished it, send to your tutor. This activity will be evaluated.

### Describing language nowadays

There are three main sources from where you can get information about a language: dictionaries, didactic materials and grammar books (when we think about the teaching/learning relation in schools, of course). Nevertheless, it's a very passive action: you open the books and consult them; if you don't understand what is written or you don't get the information you need, you need the help of a teacher or get another material.

With the Corpus Linguistics approach, these classical learning methods are being reviewed. It's time to teach your students how and where they can get information about the language in a more active way! It's time to teach them how to solve their own questions and go beyond without the help of a teacher.

Among other areas, the Corpus Linguistics approach can work with: Phonetics and Phonology; frequency of the most common words in a language; frequency of grammatical classes; Morphology and morphosyntactic variations; Syntax; collocation comparisons in a language (such as adverbials, for example); recognition of composite (as binomials) and complex (or n-grams) Lexicon; Phraseologisms (sayings, idioms); verbal and nominal regency; selection of nomenclature and database for terminological works (in all areas); general mono-or multilingual dictionaries creation; verification of translation modalities in mono-or bilingual corpora; Dialectology; translators database; foreign language teaching and development of teaching materials (ESL); literary, technical and journalistic (parallel corpora) translations review; aid for Text, Gender and Discourse studies; Prosody; creation of computerized tools (such as lexical and grammar checkers, summarization, text mining, textual simplification and elaboration); natural language processing (NLP) - automated translations; Interlanguage description in foreign language teaching; Stylistics analysis; Pragmatics; description and analysis of "errors" in written texts in native or foreign language; Semantics; human and machine translation; Metaphors study and treatment.



There are many possibilities of using Corpus Linguistics in other areas than Linguistics and Language. Learn about one here, with the text of Oliria Mendes Gimenes: <https://repositorio.ufu.br/handle/123456789/13884>

## Choosing an area to be described: Terminography

Terminology is the branch of Linguistics that studies the vocabulary of a specific area of knowledge. For example, we can study the terminology of Chemistry, the terminology of Education, etc. Terminography is the practical branch related to Terminology. When you create a specialized dictionary for an area of knowledge, you are doing a terminographical work.

Terminographical works generally, nowadays, use corpora as basis to create the microstructure of a dictionary entry. The examples we get from a corpus help us to determine the grammar structure of the word, its field, its semantical relations, etc. More important, they help us build the definition of a specialized word.

To obtain the examples, just access the corpus, using the AntConc program (as we have previously done), and go to the concordance lines: they supply information you can use in the dictionary's microstructure.

For the definition, a good idea is to seek the word you want plus the verb to be. Take a look at figure 62. It's the same of figure 61, but we added the word *is* in the search field:

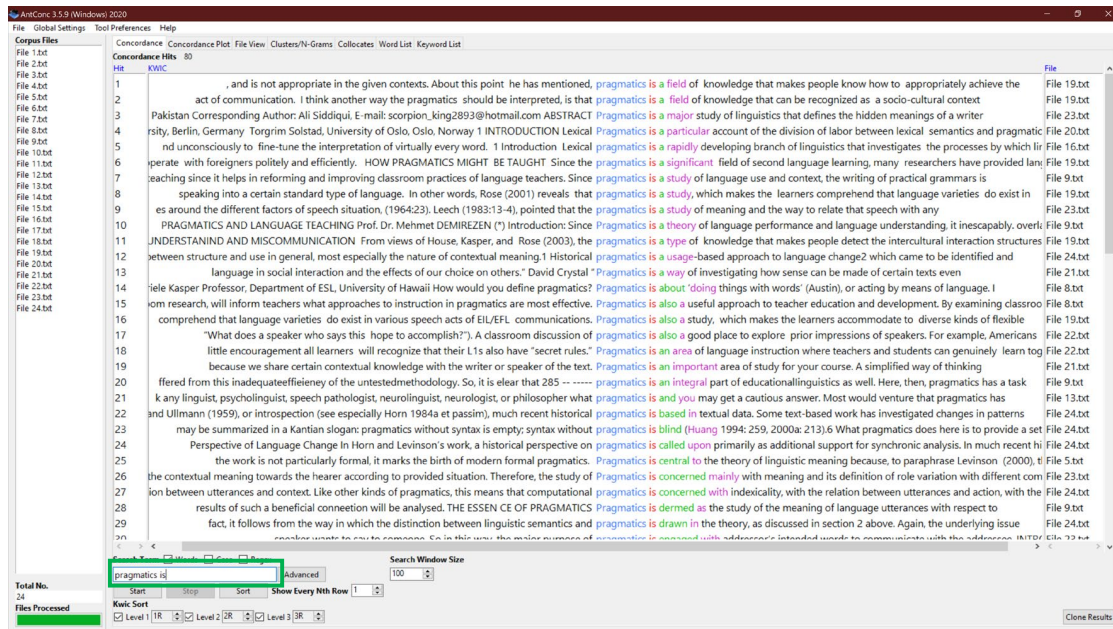


Figure 61. Concordances for the combination PRAGMATICS + IS.

The concordance lines you get are serious candidates of what Aubert (1996) calls Definitory Contexts, a good definition of the term. In this case, we have 80 concordance lines that can supply us with information about what is PRAGMATICS. Other possibilities of search to find Definitory Contexts are apposition (using commas), brackets and colon. In figure 62, we searched PRAGMATICS + colon:

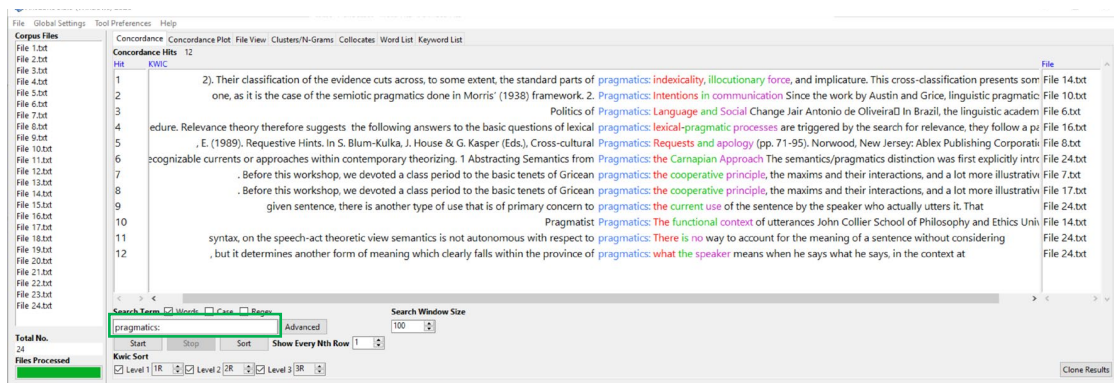


Figure 62. Concordance for the combination PRAGMATICS + :.

To organize all the information you get from the concordance lines, in terminographical works, you need a specific program. The program can be installed in your computer (for example, some researchers still use Access, from Microsoft) or you can use online solutions.

## Web environment for terminological management: VoTec

The VoTec (Vocabulário Técnico; FROMM, 2007) is an online solution for terminographical projects. Nowadays we call this type of program web environment for terminological management. You can see some examples of how the system works in here: <http://treino.votec.ileel.ufu.br/>. This is the page of the product of researches done with the VoTec<sup>14</sup>.

Online Technical Vocabulary Full Screen | Português | Help

Linguistics Choose an area

Search

Kind of exhibition  
Normal  
Extended

Kind of search  
Total  
Translator  
Modular

External Searches  
Corpus NILC  
Google  
Answers.com  
Wikipedia  
CORTEC

English

**Results from your search for: ""**

- [Allophone](#) in the area: [Linguistics](#) > [Theoretical Linguistics](#) > [Phonetics](#)
- [Derivation](#) in the area: [Linguistics](#) > [Theoretical Linguistics](#) > [Morphology](#)
- [Grammar](#) in the area: [Linguistics](#) > [Theoretical Linguistics](#) > [Syntax](#)
- [Intonation](#) in the area: [Linguistics](#) > [Theoretical Linguistics](#) > [Phonology](#)
- [Language](#) in the area: [Linguistics](#) > [Theoretical Linguistics](#) > [Syntax](#)
- [Meaning](#) in the area: [Linguistics](#) > [Theoretical Linguistics](#) > [Syntax](#)
- [Morphemes](#) in the area: [Linguistics](#) > [Theoretical Linguistics](#) > [Morphology](#)
- [Morphology](#) in the area: [Linguistics](#) > [Theoretical Linguistics](#) > [Morphology](#)
- [Phoneme](#) in the area: [Linguistics](#) > [Theoretical Linguistics](#) > [Phonology](#)
- [Phonetics](#) in the area: [Linguistics](#) > [Theoretical Linguistics](#) > [Phonetics](#)
- [Phonology](#) in the area: [Linguistics](#) > [Theoretical Linguistics](#) > [Phonology](#)
- [Speech](#) in the area: [Linguistics](#)
- [Stem](#) in the area: [Linguistics](#) > [Theoretical Linguistics](#) > [Morphology](#)
- [Vowels](#) in the area: [Linguistics](#) > [Theoretical Linguistics](#) > [Phonology](#)

Português

**Resultados da busca por: ""**

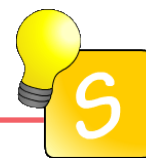
- [Afixos](#) na área: [Linguística](#) > [Linguística Teórica](#) > [Morfologia](#)
- [Alofone](#) na área: [Linguística](#) > [Linguística Teórica](#) > [Fonética](#)
- [Derivação](#) na área: [Linguística](#) > [Linguística Teórica](#) > [Morfologia](#)
- [Entonação](#) na área: [Linguística](#) > [Linguística Teórica](#) > [Fonologia](#)
- [Fonema](#) na área: [Linguística](#) > [Linguística Teórica](#)
- [Fonema](#) na área: [Linguística](#) > [Linguística Teórica](#) > [Fonologia](#)
- [Fonologia](#) na área: [Linguística](#) > [Linguística Teórica](#) > [Fonologia](#)
- [Fonologia](#) na área: [Linguística](#) > [Linguística Teórica](#) > [Fonologia](#)
- [Fonética](#) na área: [Linguística](#) > [Linguística Teórica](#) > [Fonética](#)
- [Fonética](#) na área: [Linguística](#) > [Linguística Teórica](#) > [Fonética](#)
- [Gramática](#) na área: [Linguística](#) > [Linguística Teórica](#) > [Sintaxe](#)
- [Língua](#) na área: [Linguística](#) > [Linguística Teórica](#) > [Sintaxe](#)
- [Morfemas](#) na área: [Linguística](#) > [Linguística Teórica](#) > [Morfologia](#)
- [Radical](#) na área: [Linguística](#) > [Linguística Teórica](#) > [Morfologia](#)
- [Significado](#) na área: [Linguística](#) > [Linguística Teórica](#) > [Sintaxe](#)
- [UF](#) na área: [Linguística](#)
- [Vogais](#) na área: [Linguística](#) > [Linguística Teórica](#) > [Fonologia](#)
- [morfologia](#) na área: [Linguística](#) > [Linguística Teórica](#) > [Morfologia](#)

Figure 63. Main consulting page of VoTec (in English), Linguistics area.

VoTec was developed as a bilingual (English/Portuguese) tool. If you develop a project, you must work with these two languages. As a result, all the processes we presented

The first step in VoTec is to have your name registered in the system. As soon as you're granted access to the system, you can start creating a term.

<sup>14</sup> Take a look at new possibilities with this version of VoTec, that works with TV Series: <http://ic.votec.ileel.ufu.br>



If you're interested in using VoTec for the class plan project of activity 12, get in touch with your tutor. We must register your name in the system first so you can access the data bank. In this case, you're going to work with the pages of the **process** involved in VoTec to create a specialized dictionary.

Let's see the first page:

Figure 64. VoTec researcher's first page.

The first step is to create a NEW TERM. Click the button; insert the name of the term and the areas involved<sup>15</sup>:

Figure 65. Creating the term PROSODY.

<sup>15</sup> Here, instead of the Pragmatics area, we're working with the Prosody area.



In the next screen, you get the name of the term you're going to work with and you can start inserting the concordance lines (contexts) you got from AntConc. In this case, let's insert the information from the fourth concordance line for PROSDOY + IS (figure 66):

Figure 66. Inserting a context for PROSODY.

As you can see, after inserting an extract of a concordance line from AntConc, we must present a concept for it (summarize the main idea). The font can continue OPDF and you must insert a date. After you do these steps, click the **SAVE** button.



You don't have to insert all the contexts, just the ones you consider the best to create a definition. It means you must understand the text very well and you must be capable of extracting some basic concepts from it. These basic concepts will form your definition.

As a result, we get a new screen, with the context already inserted in the data bank. In example 56, we got a screen with 3 examples



**Contextos Cadastrados**

Exemplo	Conceito	Fonte	Ações
Although these studies provide good descriptive evidence for the idea that prosody is linked to fluent and expressive reading, they do not help us understand whether prosody is helpful for comprehension in Young readers.	linked to fluent and expressive reading	PDF 20/04/2014	<a href="#">editar</a> - <a href="#">excluir</a>
Prosodic reading, or reading with expression, is considered one of the hallmarks of fluent reading. The major purpose of the study was to learn how reading prosody is related to decoding and Reading comprehension skills.	prosody is related with expression	PDF 20/04/2014	<a href="#">editar</a> - <a href="#">excluir</a>
As noted earlier, Chafe (1988) and others suggested that prosody is a reflection of comprehension.	reflection of comprehension	PDF 20/04/2014	<a href="#">editar</a> - <a href="#">excluir</a>

Contextos Cadastrados: 3

25/05/2014 07:18 © 2007 FFLCH - ICMC Jr.

Figure 67. Inserted contexts for PROSODY.

When you finish inserting the contexts, click the button **NEXT STEP** (figure 66). You get a new page with this appearance:

**Vocabulário Técnico Online** Tela Cheia | English

**Termo: Prosody**

Exemplo	Conceito	Fonte
1 Prosodic reading, or reading with expression, is considered one of the hallmarks of fluent reading. The major purpose of the study was to learn how reading prosody is related to decoding and Reading comprehension skills.	prosody is related with expression	PDF 20/04/2014
2 As noted earlier, Chafe (1988) and others suggested that prosody is a reflection of comprehension.	reflection of comprehension	PDF 20/04/2014
3 Although these studies provide good descriptive evidence for the idea that prosody	linked to fluent and expressive reading	PDF 20/04/2014

Dados

Ontologia: Linguística > Linguística Teórica > Prosódia

Categoria Gramatical: 
 Número:

Gênero: 
 Sigla/Acrônimo:

Entrada por Extenso:

Var. Morfossintáticas:

Acepção Nº:

25/05/2014 07:26 © 2007 FFLCH - ICMC Jr.

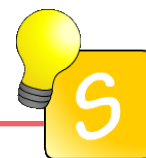
Figure 68. Second page for the term PROSODY

Each of the tabs in the middle of the page must be completed with the information from the concordance lines you have already inserted in (that is shown in the first part of the page, as you can see in figure 68). Let's see one by one:

The DATA tab must be filled with morphossyntatic and corpus information about the term you're working with:

Figure 69. Data tab.

To get the position in frequency order and the number of occurrences of the term, get back to the AntConc's Word List.



Almost hundred percent of words used in technical areas are nouns. Moreover, almost hundred percent of the English nouns have a neutral gender.

The Distinctive Traces tab must contain the concepts you prepared in the previous page. The lines represent the number of contexts you inserted in the system, the columns, the synonyms you can get from these concepts. As you can see in figure 70, we inserted the concepts expression and expressive reading in the same column (A):

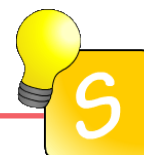
	A	B	C	D	E
1	expression				
2		comprehens...			
3	expressive...		fluency		

Figure 70. Distinctive Traces tab.

In the Semantics tab, you must get dictionary information and all the semantic traces you can get from the examples:

Figure 71. Semantics tab.

Don't worry about the definition you copy from other dictionary (DICTIONARIZED INFORMATION). They'll not be shown in the main consulting page of VoTec. We fill in the field just to help us understand the term better. In the NOTES field, you can write anything you consider important for you (also not shown in the consulting page).



Doubts about the terms you read in figure 71? Search them in the Linguistics area (see picture 54) of VoTec (you'll find the explanations in Portuguese).

We have already explained that VoTec is a bilingual platform. As you as you have inserted the terms in both languages (in this case: PROSODY and PROSÓDIA), you must link hem through the EQUIVALENT TERM tab:

Figure 72. Equivalent Term tab.

In the CROSS-REFERENCE TERMS tab, if available (it means that this term must already been inserted in the system), you can link the term you're working with terms that you consider relevant for a better comprehension of it.

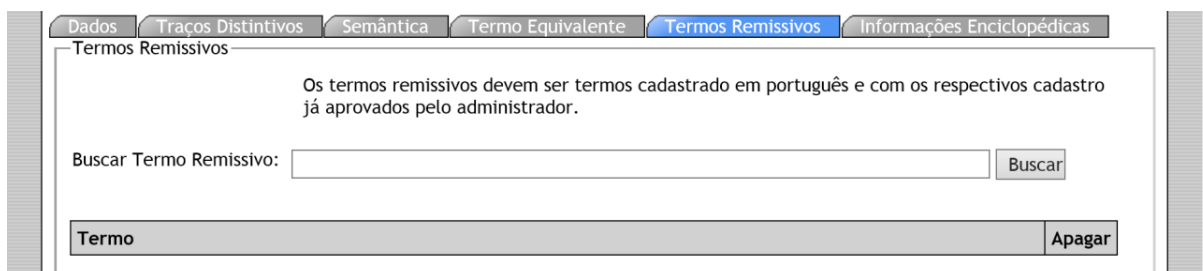


Figure 73. Cross-reference terms tab.

The information you insert in the ENCYCLOPEDIA INFORMATION tab can come from Wikipedia:

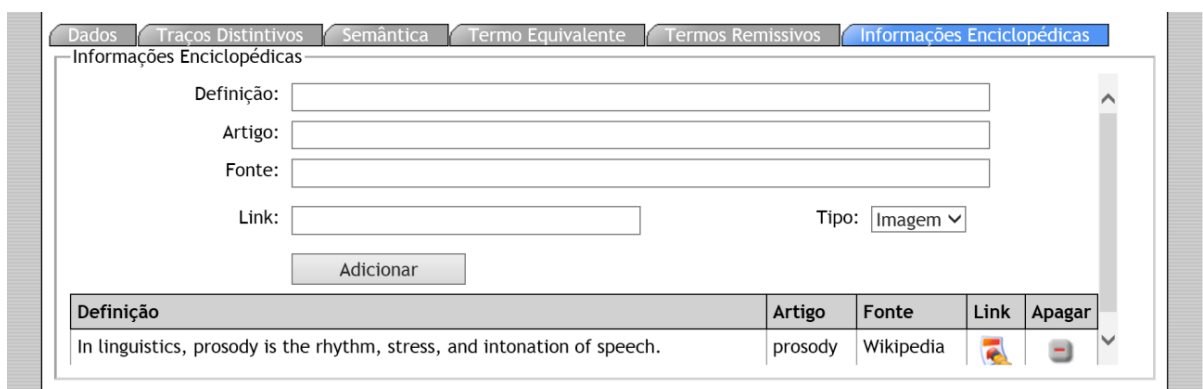


Figure 74. Encyclopedic Information tab.

The final step, using the DISTINCTIVE TRACES tab, is to fill in the FINAL CONCEPT and DEFINITION fields.

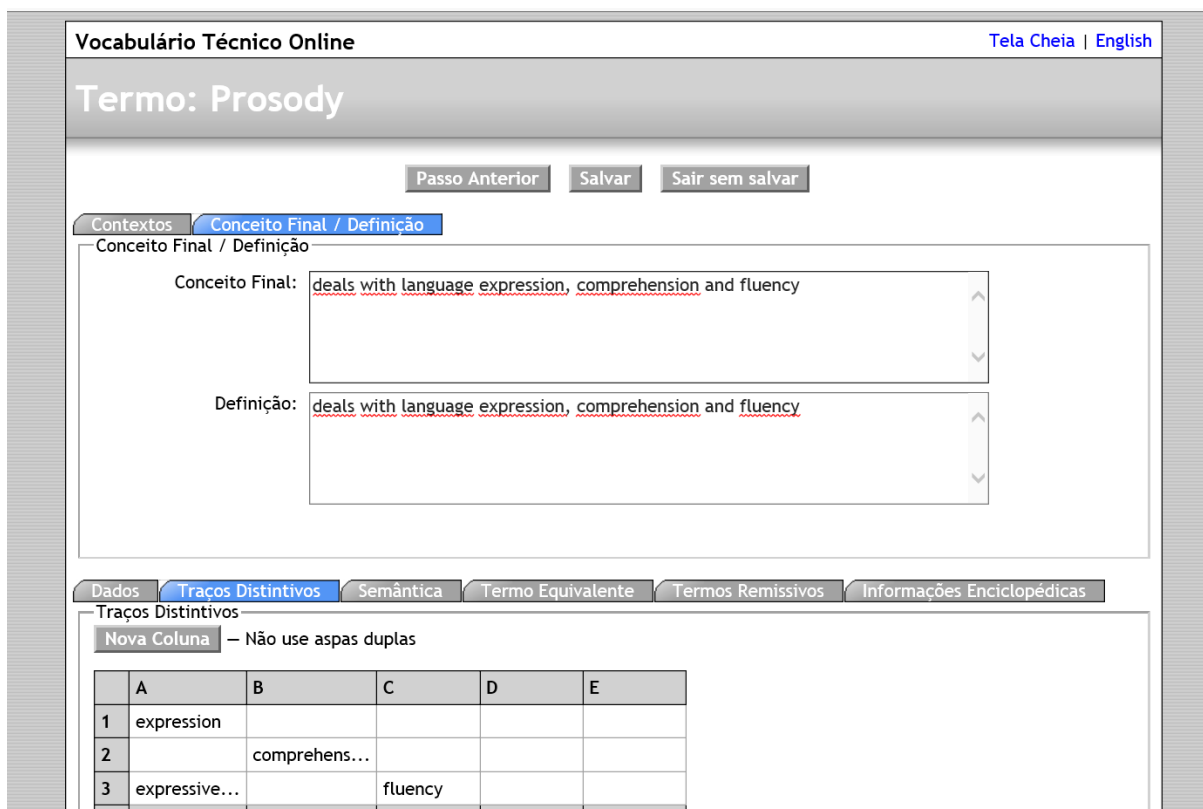


Figure 75. Final Concept and Definition tabs.

The FINAL CONCEPT tab is used to join the concepts you have in the DISTINCTIVE TRACES tab. As a definition is a sentence, you must rewrite them as a clear and synthetic sentence. After you finish these processes and the term is approved, you got a screen like this one:

Vocabulário Técnico Online Tela Cheia | English | Ajuda

Linguística Escolha uma área

Prosody Buscar

**Tipos de Exibição**  
Normal  
Descritiva

**Tipos de Consulta**  
Total  
Tradutor  
Modular

**Consultas Externas**  
Corpus NILC  
Google  
Answers.com  
Wikipedia  
CORTEC

Português
Nenhum termo encontrado equivalente a "Prosody"
English
<a href="#">Go back to search results</a>
<b>Prosody.</b> <i>Prosodics.</i> n.m/f.s. deals with language expression, comprehension and fluency. Ex.: Prosodic reading, or reading with expression, is considered one of the hallmarks of fluent reading. The major purpose of the study was to learn how reading prosody is related to decoding and Reading comprehension skills.. <i>Co-hyponyms:</i> comprehension. <i>Corpus:</i> <i>Frequency order position:</i> (4); <i>Term number of occurrences:</i> (388). <b>Encyclopedic Information:</b> In linguistics, prosody is the rhythm, stress, and intonation of speech. em: <i>prosody</i> - <a href="#">Wikipedia</a>

Figure 76. Term PROSODY in VoTec's page (just English).

To have an idea how a complete term is viewed in the system, let's take a look:

Vocabulário Técnico Online Tela Cheia | English | Ajuda

Linguística Escolha uma área

Buscar

**Tipos de Exibição**  
Normal  
Descritiva

**Tipos de Consulta**  
Total  
Tradutor  
Modular

**Consultas Externas**  
Corpus NILC  
Google  
Answers.com  
Wikipedia  
CORTEC

Português
<a href="#">voltar ao resultado da busca</a>
<b>Gramática.</b> <i>Sintaxe.</i> s.f.s. conjunto de regras internalizadas na mente dos indivíduos de uma determinada comunidade linguística; uma teoria sobre uma língua particular; deve refletir a maneira como o falante constrói enunciados. Ex.: a gramática de uma língua natural é uma teoria sobre a Língua-l de um indivíduo.. <i>hipônimo de:</i> língua. <i>Co-hipônimos:</i> regras internalizadas; teoria sobre a língua. <b>Córpus:</b> <i>Posição na Ordem de Frequência:</i> (93); <i>Nº de Ocorrências do termo:</i> (306). <b>Informações Enciclopédicas:</b> Gramática é o conjunto de regras individuais usadas para um determinado uso de uma língua, não somente da norma culta, mas também de variantes não padrão. É ramo da Linguística que tem por objetivo estudar a forma, a composição e todas as questões adicione Em: <i>Gramática</i> - <a href="#">Wikipedia</a>
English
<a href="#">Go back to search results</a>
<b>Grammar.</b> <i>Syntax.</i> <i>Grammar.</i> n.m/f.s. inventory which aims at the building of the language structure; allows mappings between meanings and signals incorporating meaning into the utterance; it is an internally structured set of rules, autonomous of meaning, shaped and reshaped through language use, responding to communicative challenges. Ex.: We describe the grammar is structured internally, and how it adds structure to utterances and decodes it again.. <i>Synonyms:</i> principles. <i>Hyponym of:</i> meaning; utterances; language; system; communication. <i>Hypernym of:</i> mappings; signals; constructions; form-meaning pairings. <b>Corpus:</b> <i>Frequency order position:</i> (3); <i>Term number of occurrences:</i> (2278). <b>Encyclopedic Information:</b> In linguistics, grammar is the set of structural rules that govern the composition of clauses, phrases, and words in any given natural language. The term refers also to the study of such rules, and this field includes morphology, syntax, and phonology, of em: <i>Grammar</i> - <a href="#">Wikipedia</a>

25/05/2014 08:30 © 2007 Guilherme Fromm - ICMC Jr.  
Termo elaborado por [Virgínia do Nascimento Peixoto](#) (pt) [Marcio Issamu Yamamoto](#) (en)

Figure 77. Term GRAMMAR in the VoTec.

VoTec has many ways to display the information from the data bank. **Try them!**

## REFERENCES



ANTHONY, L. **AntConc version 3.4.1**. Disponível em: <http://www.antlab.sci.waseda.ac.jp/index.html>. Acessado em: 20 abril 2014.

AUBERT, F. H. **Introdução à metodologia da pesquisa terminológica bilíngüe**. São Paulo: Humanitas, 1996.

BERBER SARDINHA, T. **Linguística de Corpus**. Barueri: Manole, 2004.

FROMM, G. **VoTec**: a construção de vocabulários eletrônicos para aprendizes de tradução. São Paulo, 2007. Tese (Doutorado – Programa de Pós-Graduação em Estudos Linguísticos e Literários em Inglês – Departamento de Letras Modernas). Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo.

FROMM, G; YAMAMOTO, M. I. Terminologia, terminografia, tradução e linguística de *corpus*: a criação de um vocabulário bilíngüe sobre linguística. TAGNIN, S. E. O.; BEVILACQUA, C. (org.). **Corpora na terminologia**. São Paulo: Hub Editorial, 2013.

MCGLASHAN, M. ESRC Centre for Corpus Approaches to Social Science (CASS). **Corpus**: some key terms. Lancaster: Lancaster University, 2013. Disponível em: [http://cass.lancs.ac.uk/?page\\_id=956](http://cass.lancs.ac.uk/?page_id=956). Acesso em: 26 abril 2014.

SCOTT, M. **WordSmith Tools version 6**. Liverpool: Lexical Analysis Software, 2014.

## SUGGESTED BOOKS



BIBER, D; CONRAD, S; REPPEN, R. **Corpus linguistics**: investigating language structure and use. Cambridge: Cambridge University Press, 2002.

BIDERMAN, M. T. C. (1978) **Teoria Lingüística**: teoria lexical e linguística computacional. 2. ed. São Paulo: Martins Fontes, 2001. 356 p. (Coleção Leitura e Crítica).

DELGADO, H. K., FINATTO, M. J., PERNA, C. L. (org). **Linguagens especializadas em corpora**: modos de dizer e interfaces de pesquisa [recurso eletrônico]. Porto Alegre: EDIPUCRS, 2010. Disponível em: <http://www.pucrs.br/edipucrs/linguagensespecializadasemcorpora.pdf>

MCENERY, T. **Corpus linguistics**: an introduction. Edinburgh: Edinburgh University Press, 2001.

SARDINHA, T. B. (org). **A língua portuguesa no computador**. Campinas: Mercado de Letras, 2005.

SARDINHA, T. B. **Pesquisa em linguística de Corpus com WordSmith Tools**. Campinas: Mercado de Letras, 2009.

SINCLAIR, J. **Corpus, Concordance, Collocation**. Oxford: Oxford University Press, 1991.

VIANA, V.; TAGNIN, S. E. O. (Orgs.). **Corpora no ensino de língua estrangeira**. São Paulo: Hub Editorial. 2010.

TAGNIN, S. E. O. (org). Cadernos de Tradução/UFSC/NUT. nº 9. Florianópolis: NUT, 1996. Disponível em: <http://www.periodicos.ufsc.br/index.php/traducao/issue/view/432>