Applied Language Learning

Article

Building and Cleaning Corpora for Linguistic Analysis: A Practical Guide

Ghadi Matouq
Doctoral Candidate, Applied Linguistics, University of Memphis

Hana Alqabba
Doctoral Candidate, Applied Linguistics, University of Memphis

This guide aims to make corpus building and corpus analysis feasible and practical for language instructors and/or researchers who may view building a corpus as difficult or believe that linguistic analysis requires advanced programming skills. Many avoid creating custom corpora due to these perceived barriers, instead relying on existing corpora and basic analysis tools. We present accessible instructions for corpus building, text cleaning, and linguistic analysis based on our coursework and research experience. The guide contains two parts: theoretical foundations covering corpus linguistics definition, research questions in corpus linguistics, and different types of corpora; and practical applications including corpus construction, text preparation, automated annotation, and an introduction to some types of lexical analysis. The guide demonstrates that systematic instruction makes corpus methods accessible to language teachers and novice researchers. We emphasize that hands-on practice is essential for developing corpus research skills and encourage active application of these methods to readers' research questions. We conclude with a discussion on the benefits of corpus analysis for the language classroom.

Keywords: *Corpus Linguistics, Corpus Building, Text Cleaning, Linguistic Analysis, Corpus-Based Research*

INTRODUCTION

Corpus analysis is a valuable skillset for language instructors and researchers, as it provides opportunities for teachers to better understand language features and use (particularly related

APPLIED LANGUAGE LEARNING (ISSN 2164-0912) is the journal of the Defense Language Institute Foreign Language Center, in the U.S. Department of Defense. The views expressed herein are those of the author(s), not the Department of Defense or its elements. Further reproduction is not advisable. Whenever copyrighted materials are reproduced in this publication, copyright release has ordinarily been obtained only for use in this specific issue. Requests for reprints should be directed to the individual authors.

to vocabulary and grammar) and also design effective materials for classroom instruction or even language assessments. Yet, many language instructors or even researchers view corpus building as an overly difficult task and text analysis as a highly technical process requiring advanced programming skills. This, however, is not the case. What corpus building and corpus analysis do require is systematic thinking, as well as learning the steps and application options.

This article demonstrates how to build, clean, and analyze a corpus through step-by-step instructions, allowing language instructors and novice researchers to approach these processes with clarity, efficiency, and effectiveness. There are many ways to address the processes and decisions introduced here, but we hope that this practical guide is enough to get interested teachers and researchers started.

This guide is organized into two main parts. The first briefly provides a foundation by defining corpus linguistics, examining research questions in corpus linguistics, and exploring different types of corpora. The second section focuses on practical applications, covering corpus building, importance of text preparation and cleaning, linguistic annotation, and some types of lexical analysis. Corpus building and analysis require hands-on practice to master. Therefore, we encourage readers to follow the instructions in this guide and use the tools to build, clean, and analyze their own corpora, while also exploring other tools and processes that may work for them.

Corpora and Corpus Linguistics

A Definition

As a field of study, corpus linguistics involves studying language in use through a corpus or corpora (plural of corpus). A corpus is a collection of texts stored in digital format for linguistic research. To analyze these texts, corpus linguists use software tools that help them find patterns in the data. For example, they can examine how often a word appears, what words tend to co-occur with other words, or how words or phrases are used, showing every occurrence of a word or a phrase in context. These are just a few examples, as researchers in corpus linguistics are diverse in interests, diverse in aims, and notably creative.

Research Questions in Corpus Linguistics

Corpus linguistics serves as a research methodology that allows researchers to examine different areas of applied linguistics including grammar, lexis, discourse, pragmatics, and second language acquisition. Researchers can address both quantitative and qualitative questions depending on their research goals (Timmis, 2015).

Quantitative corpus analysis focuses on frequency information that can be generated easily from corpus data. Basic frequency questions include: What are the most frequent words in a corpus? How many instances of a given word appear? What percentage of total tokens does a word represent? What are the most frequent collocations or phrases? What are the most frequent grammatical structures? These questions can be applied with specific focus, such as examining word frequency in different domains (e.g., medical vs. legal texts), demographics (e.g., age or proficiency groups), or modes (e.g., spoken vs. written language).

However, many research questions require qualitative analysis that cannot be answered through automatic corpus analysis alone. For example, determining which meaning of a word like "tip" is most frequent, or identifying when "marvelous" is used sarcastically, requires manual examination of corpus data. Thus, it is important for researchers to read individual instances and interpret their meaning in context. This combination of quantitative patterns and qualitative interpretation makes corpus linguistics an effective methodology for understanding authentic language use across diverse research contexts.

Different Types of Corpora

Texts in corpora can be written, spoken, and multimodal. Written corpora could include books, research articles, newspapers, emails, and blogs, digitized and stored in text files. Some written texts may need initial preparation before the cleaning process begins. For example, scanned and hand-written texts require conversion into machine-readable formats before they can be used reliably. Nevertheless, written corpora are widely used in research as they are often readily available in digital form.

Unlike written corpora, spoken corpora are somewhat less common because they are harder to collect, require transcription, and are notable less 'formalized', as they are comprised of human speech. These include transcripts of conversations, narratives, lectures, debates, and conference talks. The least common type of corpora are those including multimodal components collected through the video recording. These include verbal and non-verbal data, including speech and body language (Foster & Oberlander, 2007). Multimodal corpora represent a development from spoken corpora, offering additional modes of communication including gesture, facial expression, and intonation, synchronized with transcripts. Multimodal corpora expand the kinds of communicative phenomena that can be examined, depending on the research focus (McCarthy & O'Keeffe, 2010).

Corpora can also be categorized in terms of their content, with two main broad categories: general and specialized corpora. General corpora contain a large collection of text from many different sources, genres, and domains, representing the whole language in all its diversity. Therefore, they may include texts coming from newspapers, books, academic writing, conversation, TV shows, fiction, non-fiction, etc. Specialized corpora, on the other hand, are not as large as general corpora, and their size could vary based on the language they represent. They include texts that meet specific criteria, focusing on specific registers and genres.

Several influential corpora have become classic references in corpus research, which will be used here to exemplify the distinction between general and specialized as well as written and spoken corpora through their content and mode. The Corpus of Contemporary American English (COCA) and the British National Corpus (BNC) are two important general corpora in English that are predominantly comprised of written texts, with very few spoken texts (20% of COCA and 10% of BNC come from spoken language). The TenTen family of corpora, 10+ billion word corpora available in over 50 languages such as Spanish, Japanese, Russian, Arabic, and Greek, represent large-scale web-based general corpora (Jakubíček et al., 2013). The British Academic Written English (BAWE) corpus and the Michigan Corpus of Upper-Level Student Papers (MICUSP) are written corpora representing language use in the academic domain. The British Academic Spoken English (BASE) and the Michigan Corpus of Spoken Academic English (MICASE) are spoken corpora. Both are also examples of specialized corpora representing language use in English academic domains. These existing corpora offer significant savings in time and effort, often include detailed metadata useful for coding and analysis, and can be quite large and therefore enable reliable frequency counts of linguistic features and more diverse word and phrase types than smaller corpora could provide.

However, a significant limitation of utilizing existing corpora is that they may not align perfectly with research questions and target population. The reality is that “no corpus is one size fits all” (Egbert et al., 2020, p. 5), which means that the texts included in any given corpus shape the linguistic population to which the results can be generalized. Therefore, it becomes essential to choose a corpus that matches research needs in two keyways: (1) the types of texts it includes and (2) whether it contains enough data for reliable analysis. Additionally, it is important to identify where mismatches exist and interpret findings within these limitations (Egbert et al., 2020). Further, in language for specific purposes contexts where educators frame language teaching around specific genre competencies, it is likely that publicly available specialized corpora do not exist.

PRACTICAL CONSIDERATIONS

Corpus Building

Researchers may want to compile their own corpus if existing corpora cannot be used to answer their questions. For instance, they may want to focus on a particular author, genre, or variety of language that has not been represented in available corpora. They also may build new corpora to analyze recent language use. If you decide to build your corpus, here is a step-by-step process for corpus compilation. A summarized outline of these steps is provided in Figure 1.

1. **Make a research plan.**
   Start with planning and defining research or teaching goals: begin by asking your research questions, determining the type of language that could address your research questions, and identifying specific linguistic features of interest. For example, a question about differences in the use of reporting verbs (e.g., state, claim, emphasize, discuss) between experts’ research articles and students’ MA theses. This question could be answered by collecting research articles from reputable journals and MA theses available online.

2. **Design your corpus.**
   Now that you have identified the corpus that answers your research questions, you need to design your corpus.
   a. *Set some criteria to start compiling your corpus.* Corpus compilers use a range of criteria when selecting texts for a corpus to ensure that the resulting dataset is representative, balanced, and suitable for the intended research. Among the commonly used criteria, as described by Sinclair (2004), are mode (spoken/written), type (books/research articles), domain (academic/popular), variety (specific language/languages/language varieties), location (US/UK), and date.
   b. *Determine the corpus size.* While there is no definite answer for how large a corpus should be, Reppen (2022) explains that the size of a corpus could be determined by two main factors: representativeness (ensuring sufficient data to accurately reflect the language being studied) and practicality (considering time constraints). In simple words, you need to balance the ideal with what is feasible. For highly specific research, e.g., the language of a particular author, such as 'the Swale's corpus' compiled by Hyland (2008), a more focused corpus can be fully representative of the language. But for broader studies, e.g., scrutinizing academic language in several disciplines (Hyland, 1999), texts were selected to represent the language under investigation.
   c. *Balance your corpus.* As Sinclair (2004) indicated, balance does not require accuracy in a strict sense. Instead, it demands transparent, reasoned decisions about text proportions, guided by the corpus's goals and the diversity of language. To do so, aim for a balance between different text types and document your inclusion/exclusion decisions. For example, while the text types included in the COCA corpus may not fully capture the range of public discourses in contemporary English, their diversity allows for strong insights.

3. **Collect and prepare your data.**
   There are several cases where researchers should seek and obtain permission before compiling their corpus. For example, approval from an institution's IRB is required for any human subjects research at US institutions. Also, there are cases where researchers need permission to use texts with copyright (Reppen, 2022). In any research, ethical considerations should be respected. You can start your data collection with the required permission, if any.
   a. *Develop protocols for consistent data collection*. Guidelines and detailed procedures are essential to collect reliable data, especially when multiple researchers are involved. For example, you need to decide on what to keep and what to delete from a document in your written corpus (e.g., will you keep references and tables in an academic text or bylines in a newspaper article?). Also, before transcribing texts from your spoken corpus, you need to have a transcription system and decide on how to represent non-verbal data such as laughter, pauses and overlaps, if they will be documented (Reppen & Simpson, 2002). Consider the protocols as a go-to resource for yourself, teams, and the research community. If they need to be changed, then all previously prepared data must be carefully reviewed.
   b. *Create a metadata scheme*. You need to describe your corpus information and keep your data organized and searchable. Think about what information might be useful for

analysis. Information such as context, speakers’ identity, gender, age or author, year of publication, and title are some common details, but other information may be of interest to you. Focus on what you need to understand and organize your data clearly, as well as what you may wish to analyze later.

c. *Metadata can be stored inside the text, in the filename, or in a separate file*. To contain information within the text, use headers at the beginning of each file with information about it (Reppen, 2022). You should add headers clearly and consistently. A common style is to use angle brackets < > for each line. Add a clear separator between the header and the main text, like this: <end_header>. Using brackets and separators makes it easy for you and the software to recognize and ignore the header during analysis. Metadata can also be stored separately in a CSV table (e.g., Excel) with each row for one text and each column including metadata information (e.g., title, author, year of publication).

4. **Clean the texts.**

Corpus cleaning is an essential step in preparing data for analysis.

a. Work with plain text files and save files in UTF-8 format to ensure compatibility with most corpus analysis tools (Lu, 2014; Reppen, 2022).

b. In a spoken corpus, transcripts should be cleaned by correcting human or machine transcription errors, and identifying and correcting inconsistencies in transcription, such as the use of contractions (e.g., whether to use ‘*gonna, going to, or both’*, ‘*cannot or can’t*’ following a previously established criteria), and the presentation of non-verbal elements, ensuring consistency across the dataset (Thompson, 2004).

c. In a written corpus, this involves removing undesirable components from your corpus. This includes headers, footers, page numbers, tables, charts, line breaks, extra white spaces, and copyright notices, unless those things are of interest, and other peculiarities of genres and texts of interest. Also, careful attention should be paid to character changes in letters, punctuation, or symbols. If you convert a PDF to a text file and notice that some letters look strange, it could mean the original PDF uses unusual fonts or encoding, and the converter could not read them correctly; in that case, trying a different converter or using an Optical Character Recognition (OCR) tool might help fix the issue. It is necessary to do manual checking and editing to correct errors that could result from converting the original file or using OCR. Some PDFs that use multiple columns may create similar issues, and attempting to access HTML files or other formats could save considerable time.

d. Note that the analytical focus often influences how the text should be cleaned. Sentence boundaries should be clear and accurate, especially when conducting syntactic analysis and using the sentence as the unit of analysis (e.g., measuring the number of words or clauses per sentence, using tools like TAASSC for syntactic complexity). For example, when cleaning texts with bullet points, the researchers must decide if they will add periods to bullet points that are complete sentences and add semicolons to bullet points that are fragments. Accurate sentence boundary detection is essential for accurate POS tagging (Lu, 2014), which means that sentence boundaries are also important for lexical analyses.

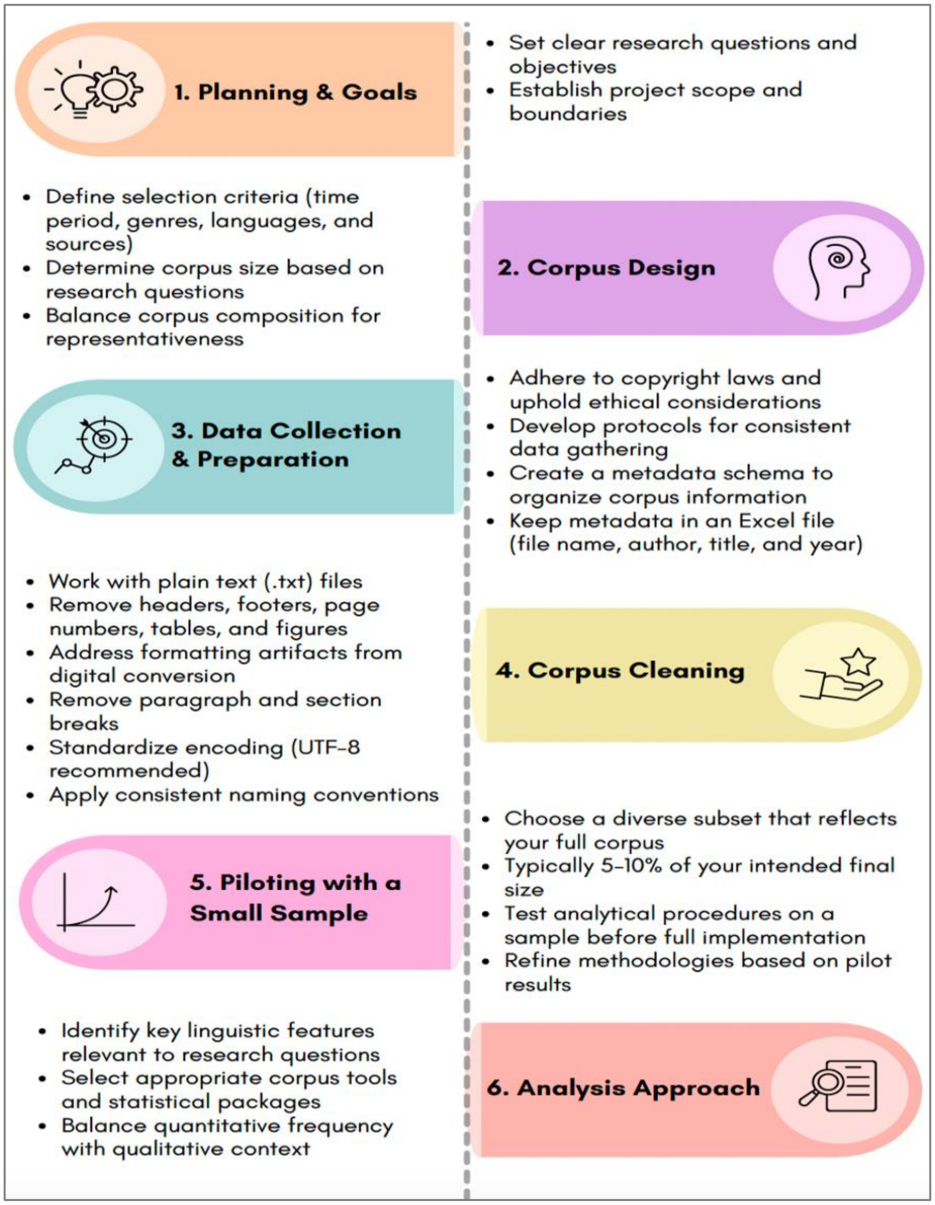
e. Apply consistent name conventions when saving your text files and the original files. Aim for short names with letters and numbers and replace spaces with underscores, as in Lec_1, RA_1 for (lecture 1) and (research article 1). You can include a clear abbreviation for other information, for example, if you have multiple disciplines, you can use CS for Computer Science, BA for Business Administration, ENG for English, and PSY for Psychology.

5. **Conduct a pilot.**
   Before cleaning the full corpus, it is good practice to conduct a pilot using a small, sample, typically around 5-10% of the total data. This subset should reflect the diversity of the full dataset in terms of genre, structure, or source. A pilot can be useful for testing cleaning procedures, such as removing noise, correcting encoding issues, or standardizing formatting. Applying cleaning methods to a sample allows researchers to identify and address potential issues early, including inconsistent spacing, problematic characters, or formatting irregularities. As part of this step, compare the original uncleaned texts to the cleaned text samples, and make sure that the arrangement of the text and paragraphs has not been changed by the conversion process. Based on the findings from the pilot, the cleaning approach can be refined to ensure it is effective. This step improves the overall quality and consistency of the corpus and helps prevent complications during later stages of analysis. See the Appendix for notes on the importance of text preparation and cleaning.

6. **Begin analysis.**
   After ensuring that the corpus is cleaned, the analysis phase can begin. It is important to identify key linguistic features related to the research questions, such as specific words, grammatical patterns, phrase structures, or discourse markers. These features guide the focus of the analysis and help ensure that data interpretation aligns with the study's objectives. Once the relevant features are identified, researchers must select appropriate corpus tools and statistical packages to support their analysis. This may include concordance lines, POS taggers, keywords, or statistical software like SPSS, depending on the complexity of the data and the nature of the analysis. Throughout the process, many researchers seek to balance quantitative findings such as frequency counts or collocation patterns with qualitative insights, including the examination of concordance lines or contextual usage. This approach strengthens the interpretation of the results and allows for a more nuanced understanding of linguistic phenomena in context.

Figure 1

Corpus Compilation Steps



1. Planning & Goals

- Set clear research questions and objectives
- Establish project scope and boundaries

2. Corpus Design

- Define selection criteria (time period, genres, languages, and sources)
- Determine corpus size based on research questions
- Balance corpus composition for representativeness

3. Data Collection & Preparation

- Adhere to copyright laws and uphold ethical considerations
- Develop protocols for consistent data gathering
- Create a metadata schema to organize corpus information
- Keep metadata in an Excel file (file name, author, title, and year)

4. Corpus Cleaning

- Work with plain text (.txt) files
- Remove headers, footers, page numbers, tables, and figures
- Address formatting artifacts from digital conversion
- Remove paragraph and section breaks
- Standardize encoding (UTF-8 recommended)
- Apply consistent naming conventions

5. Piloting with a Small Sample

- Choose a diverse subset that reflects your full corpus
- Typically 5–10% of your intended final size
- Test analytical procedures on a sample before full implementation
- Refine methodologies based on pilot results

6. Analysis Approach

- Identify key linguistic features relevant to research questions
- Select appropriate corpus tools and statistical packages
- Balance quantitative frequency with qualitative context

Linguistic Annotation of Corpus Data

After building and cleaning your corpus, begin by adding linguistic information to your texts through *linguistic annotation* if your analysis requires it. This step can be skipped if your research focuses on analyses that can be done without annotation, such as certain frequency analyses, keyword analysis, etc. However, your analysis requires linguistic annotation (such as part-of-speech tagging). You can then conduct various types of linguistic analysis including *lexical analysis* (examining word-level patterns), *syntactic analysis* (investigating grammatical structures), and more. The type of analysis you choose depends on the research questions you are trying to answer, and there are more types of annotation than those discussed here.

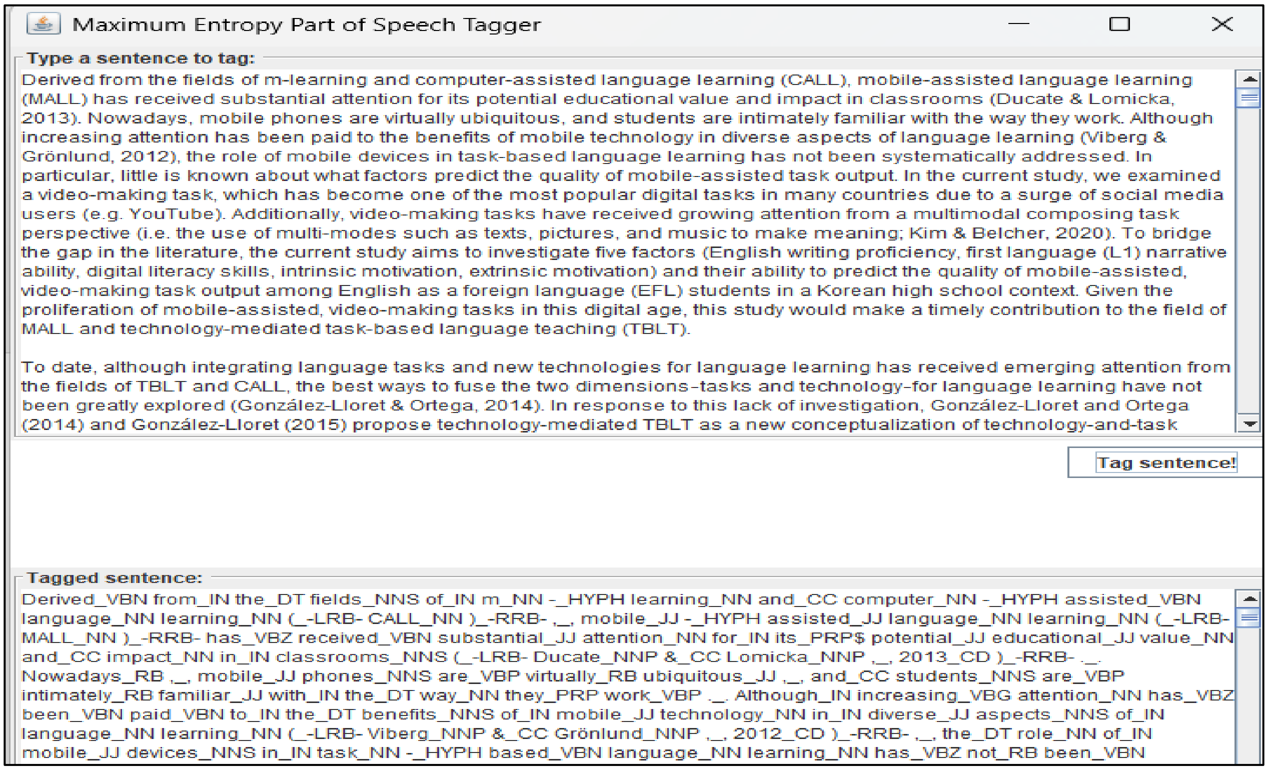
Part-of-Speech Tagging

Part-of-speech (POS) tagging is a widely used type of linguistic annotation that assigns labels or "tags" to each word in your corpus with its grammatical category such as noun, verb, or adjective (Newman & Cox, 2020). The Stanford POS Tagger is highly accurate and provides a user-friendly interface that makes it accessible for beginners in corpus linguistics, but there are other systems. You can use the tagged data as a foundation for investigating grammatical patterns, syntactic complexity, and linguistic variation in your corpus. The Stanford POS Tagger uses the Penn Treebank tagset. The tagger can also be run using Command Line prompts, but this introduces the Graphic User Interface method for novices. This tool works with English, Arabic, Chinese, French, Spanish, and German.

To conduct POS tagging with the Stanford tagger:

1. Go to the Stanford Natural Language Processing Group website and download the Stanford POS tagger (https://nlp.stanford.edu/software/tagger.shtml)
2. Ensure Java 8+ is installed on your computer
3. Extract the downloaded tagger files to a folder on your computer
4. Navigate to the tagger folder and launch the graphical interface application
5. Enter your text in the upper input window (Figure 2)
6. Click the "Tag sentence!" button
7. View the tagged results in the lower window, where each word appears with its grammatical label
8. Save or copy the tagged output in plain text for further analysis

Figure 2
Stanford POS Tagger Graphical User Interface



Derived from the fields of m-learning and computer-assisted language learning (CALL), mobile-assisted language learning (MALL) has received substantial attention for its potential educational value and impact in classrooms (Ducate & Lomicka, 2013). Nowadays, mobile phones are virtually ubiquitous, and students are intimately familiar with the way they work. Although increasing attention has been paid to the benefits of mobile technology in diverse aspects of language learning (Viberg & Grönlund, 2012), the role of mobile devices in task-based language learning has not been systematically addressed. In particular, little is known about what factors predict the quality of mobile-assisted task output. In the current study, we examined a video-making task, which has become one of the most popular digital tasks in many countries due to a surge of social media users (e.g. YouTube). Additionally, video-making tasks have received growing attention from a multimodal composing task perspective (i.e. the use of multi-modes such as texts, pictures, and music to make meaning; Kim & Belcher, 2020). To bridge the gap in the literature, the current study aims to investigate five factors (English writing proficiency, first language (L1) narrative ability, digital literacy skills, intrinsic motivation, extrinsic motivation) and their ability to predict the quality of mobile-assisted, video-making task output among English as a foreign language (EFL) students in a Korean high school context. Given the proliferation of mobile-assisted, video-making tasks in this digital age, this study would make a timely contribution to the field of MALL and technology-mediated task-based language teaching (TBLT).

To date, although integrating language tasks and new technologies for language learning has received emerging attention from the fields of TBLT and CALL, the best ways to fuse the two dimensions-tasks and technology-for language learning have not been greatly explored (González-Lloret & Ortega, 2014). In response to this lack of investigation, González-Lloret and Ortega (2014) and González-Lloret (2015) propose technology-mediated TBLT as a new conceptualization of technology-and-task

```
Derived_VBN from_IN the_DT fields_NNS of_IN m_NN -_HYPH learning_NN and_CC computer_NN -_HYPH assisted_VBN
language_NN learning_NN (_-LRB- CALL_NN )_-RRB- ,_, mobile_JJ -_HYPH assisted_JJ language_NN learning_NN (_-LRB-
MALL_NN )_-RRB- has_VBZ received_VBN substantial_JJ attention_NN for_IN its_PRP$ potential_JJ educational_JJ value_NN
and_CC impact_NN in_IN classrooms_NNS (_-LRB- Ducate_NNP &_CC Lomicka_NNP ,_, 2013_CD )_-RRB- ._.
Nowadays_RB ,_, mobile_JJ phones_NNS are_VBP virtually_RB ubiquitous_JJ ,_, and_CC students_NNS are_VBP
intimately_RB familiar_JJ with_IN the_DT way_NN they_PRP work_VBP ._. Although_IN increasing_VBG attention_NN has_VBZ
been_VBN paid_VBN to_IN the_DT benefits_NNS of_IN mobile_JJ technology_NN in_IN diverse_JJ aspects_NNS of_IN
language_NN learning_NN (_-LRB- Viberg_NNP &_CC Grönlund_NNP ,_, 2012_CD )_-RRB- ,_, the_DT role_NN of_IN
mobile_JJ devices_NNS in_IN task_NN -_HYPH based_VBN language_NN learning_NN has_VBZ not_RB been_VBN
```

Lexical Analysis: An Introduction

Texts that have been cleaned can be effectively analyzed using corpus tools to answer research queries. Lexical analysis is an essential area of corpus linguistics that has been associated with the development of corpus linguistics and Sinclair's contribution to the field through the COBUILD (Collins Birmingham University International Language Database) project. Lexical analysis examines the frequency, use, and type of words and phraseological patterns within a text collection. The following subsections provide step-by-step instructions for conducting various types of lexical analysis using specialized corpus tools. Also discussed is corpus analysis with AntConc (Anthony, 2023), a free and feature-rich tool, including an overview of its functionalities and affordances, and a phrase-frame analysis using KfNgram. Together, these sections demonstrate how to analyze different linguistic features, from individual word frequencies and collocations to complex discontinuous multi-word patterns.

Corpus Analysis with AntConc

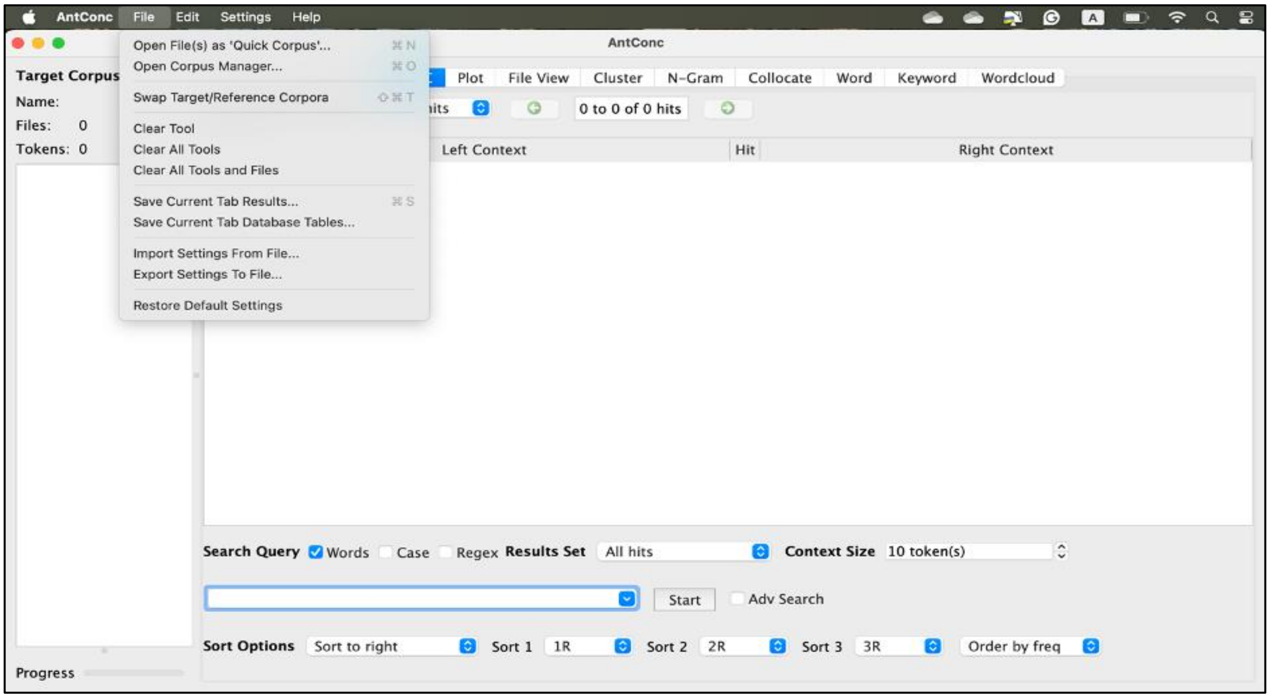
AntConc, developed by Laurence Anthony, is a widely used, freely available tool for corpus analysis, compatible with Windows, macOS, and Linux, that can be downloaded from here https://www.laurenceanthony.net/software/antpconc/.

Before beginning any analysis, you need to do the following (see Figure 3):

1. Download and install AntConc from the official website
2. Prepare your corpus files in plain text (.txt) format, although it does accept other file forms
3. Organize your files in a folder
4. Load your corpus: open AntConc → click File → open corpus manager → corpus source → raw file(s) → name your corpus → add files from your corpus folder → click on create

Figure 3

AntConc: Corpus Loading

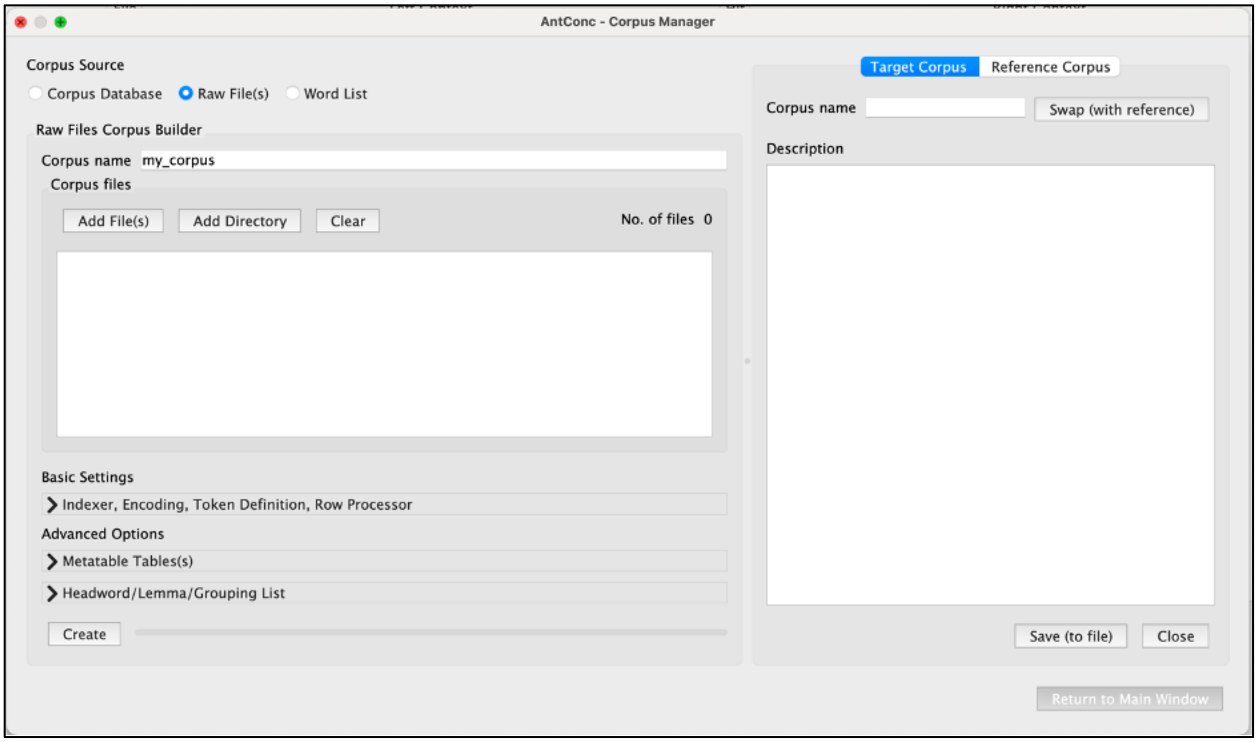

AntConc
File
Edit
Settings
Help
Open File(s) as 'Quick Corpus'... ⌘N
Open Corpus Manager... ⌘O
Swap Target/Reference Corpora ⇧⌘T
Clear Tool
Clear All Tools
Clear All Tools and Files
Save Current Tab Results... ⌘S
Save Current Tab Database Tables...
Import Settings From File...
Export Settings To File...
Restore Default Settings
AntConc
Target Corpus
Plot
File View
Cluster
N-Gram
Collocate
Word
Keyword
Wordcloud
Left Context
Hit
Right Context
Start
Progress

Name:
Files: 0
Tokens: 0
0 to 0 of 0 hits
Search Query ☑ Words □ Case □ Regex
Results Set: All hits
Context Size: 10 token(s)
□ Adv Search
Sort Options: Sort to right
Sort 1: 1R
Sort 2: 2R
Sort 3: 3R
Order by freq

Under File, you have two options to load a corpus: Open File(s) as a Quick Corpus or Open Corpus Manager. Use Quick Corpus File(s) to open text files for immediate analysis, but use Corpus Manager if you want to do repeated analyses as you can access your corpus later from the current database (any corpus you create using Corpus Manager will be listed there); or if you need to do comparative analysis using target and reference corpora as will be explained in the keyword analysis (see Figure 4).

Figure 4

AntConc: Corpus Manager



AntConc - Corpus Manager

Corpus Source

○ Corpus Database ● Raw File(s) ○ Word List

Raw Files Corpus Builder

Corpus name: my_corpus

Corpus files

No. of files: 0

Basic Settings

Advanced Options

Target Corpus

Corpus name:

Description

Before discussing analytical approaches, it is essential first to explain the functions of the interface tools in AntConc. More information about these tools and the menu options can be found at https://www.laurenceanthony.net/software/antconc/releases/AntConc424/help.pdf

KWIC (Key Word in Context): This tool shows search words in their immediate context, resulting in concordance lines with your search word in the center. You can use KWIC to examine how words are used in context, study patterns of usage and meaning, and conduct qualitative analysis of language use.

Plot: This tool visualizes the distribution of search words across texts. It shows search results as barcodes, where each bar represents a text. The text length is normalized to fit the bar width, and a vertical line on the bar shows the hits. You can use plot to see if keywords appear in specific sections, compare distribution patterns across different texts, and identify topics that appear throughout vs. in specific sections.

File View: This tool displays full text files with search words highlighted. You can use File View to understand broader discourse patterns, examine document structure, and compare and verify concordance findings in a fuller context.

Cluster: This tool identifies word combinations that appear around a specific search word, resulting in lists of word clusters around your search word. You can use Cluster to study the behavior of particular terms and do targeted lexical analysis.
N-Gram: This tool displays contiguous word sequences in the entire corpus (lexical bundles). It results in lists of chunks with frequency data. You can use the N-Gram to identify lexical bundles, examine their frequency, and examine their distribution.

Collocate: This tool finds words that frequently co-occur with search words, showing statistical associations between words appearing within a defined span. You can use this tool to conduct collocation analysis, do semantic prosody analysis (positive/negative associations), identify semantic fields and domains, and critically analyze word choices in particular contexts.

Word: This tool generates frequency lists of all words in a corpus, providing a list of words ranked by frequency, range, and percentage. You can use this tool to identify high-frequency and low-frequency vocabulary, compare vocabulary across different corpora, and study lexical density and complexity.

Keyword: This tool can be used to compare a target corpus against a reference corpus and show statistically significant words that distinguish your corpus. You can use this tool to identify distinctive vocabulary and compare corpora of different genres or periods.

Wordcloud: This tool provides a visual representation of word frequencies. You can use this tool for creating word clouds of words in your corpus by selecting the source of the data (e.g., Word, Keyword), or by inserting a text in the scratchpad.

Frequency Analysis

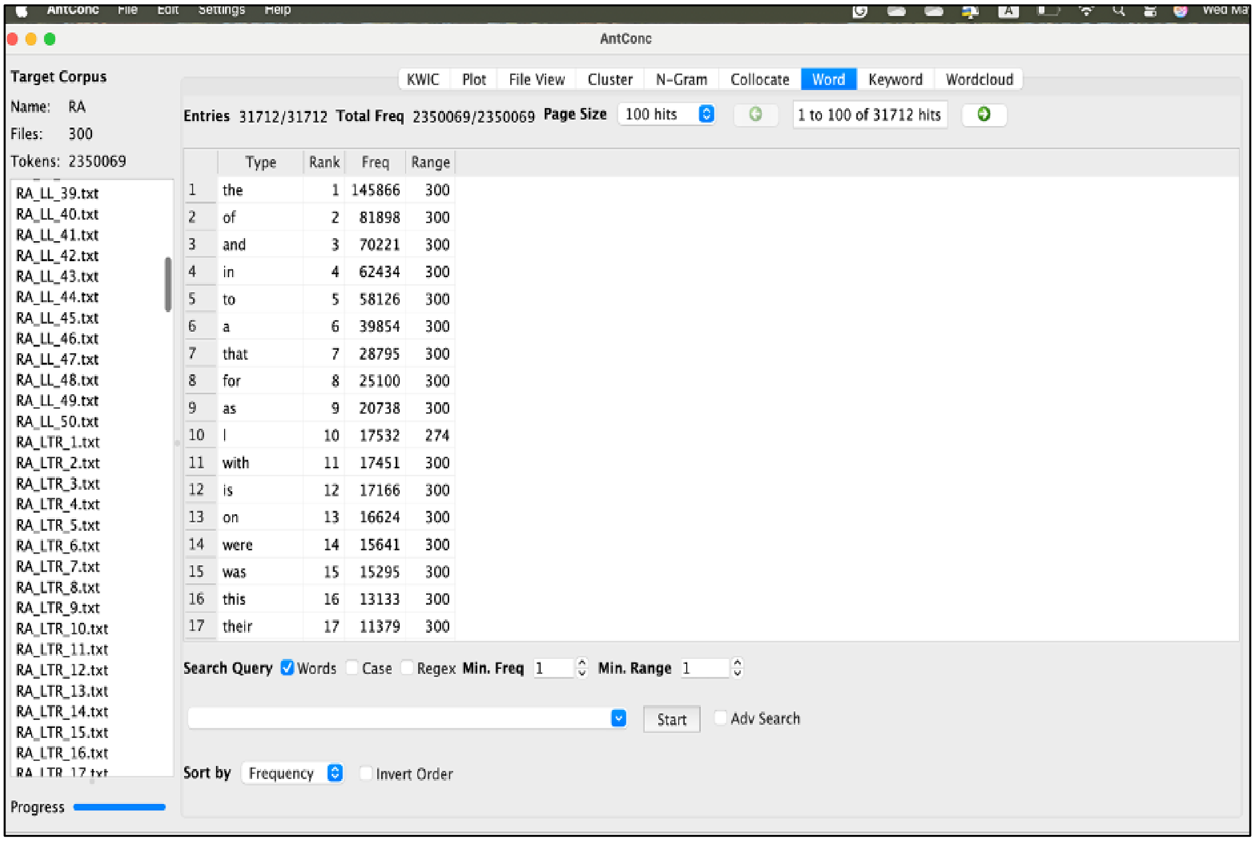
Frequency analysis reveals which words appear most frequently in your corpus, providing insights into words more common or even specific to the corpus.

To start a frequency analysis:

1. Load your corpus into AntConc
2. Click the Word tab (Figure 5)
3. Set the options for the search, minimum range, and frequency
4. Click Start to generate the frequency list
5. Selecting “Frequency” as the sort option shows the results from most to least frequent. Choosing “Range” arranges results based on how widely the words occur across texts.

Figure 5

AntConc: Word



Target Corpus
Name: RA
Files: 300
Tokens: 2350069

RA_LL_39.txt
RA_LL_40.txt
RA_LL_41.txt
RA_LL_42.txt
RA_LL_43.txt
RA_LL_44.txt
RA_LL_45.txt
RA_LL_46.txt
RA_LL_47.txt
RA_LL_48.txt
RA_LL_49.txt
RA_LL_50.txt
RA_LTR_1.txt
RA_LTR_2.txt
RA_LTR_3.txt
RA_LTR_4.txt
RA_LTR_5.txt
RA_LTR_6.txt
RA_LTR_7.txt
RA_LTR_8.txt
RA_LTR_9.txt
RA_LTR_10.txt
RA_LTR_11.txt
RA_LTR_12.txt
RA_LTR_13.txt
RA_LTR_14.txt
RA_LTR_15.txt
RA_LTR_16.txt
RA_LTR_17.txt

Entries 31712/31712 Total Freq 2350069/2350069 Page Size 100 hits 1 to 100 of 31712 hits

	Type	Rank	Freq	Range
1	the	1	145866	300
2	of	2	81898	300
3	and	3	70221	300
4	in	4	62434	300
5	to	5	58126	300
6	a	6	39854	300
7	that	7	28795	300
8	for	8	25100	300
9	as	9	20738	300
10	I	10	17532	274
11	with	11	17451	300
12	is	12	17166	300
13	on	13	16624	300
14	were	14	15641	300
15	was	15	15295	300
16	this	16	13133	300
17	their	17	11379	300

Search Query ☑ Words □ Case □ Regex Min. Freq 1 Min. Range 1

□ Adv Search

Sort by Frequency □ Invert Order

In this image, you can see the corpus information: name, files, and tokens. Word, the third tap from the last one, is selected. You can specify your search by choosing the options next to the search query. You will probably notice that function words, including prepositions and conjunctions, always ranked as the topmost frequent in the list. Researchers are most likely interested in content words rather than function words. Several studies used frequency analysis to identify domain-specific words and compare word patterns across different text types.

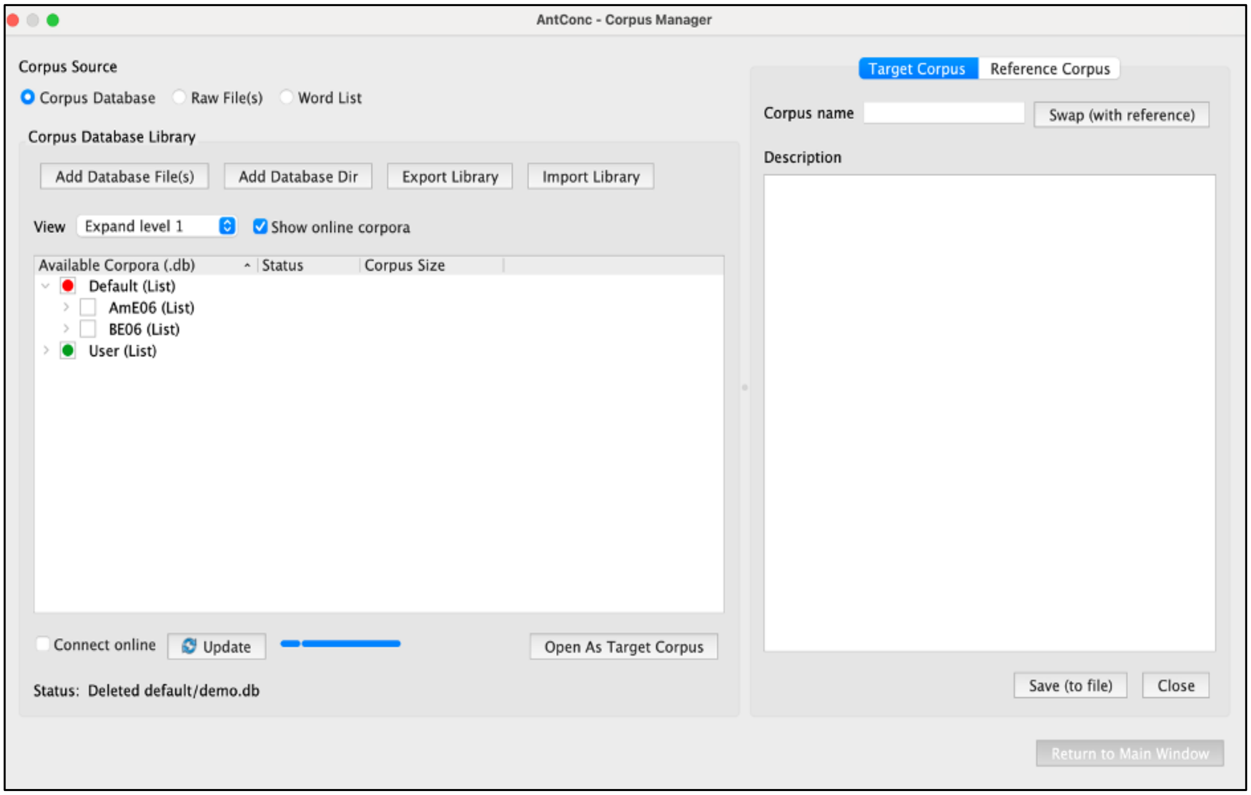
Keyword Analysis

Keyword analysis shows words occurring more frequently in a corpus than in another. Hence, it involves using two corpora: the target corpus and the reference corpus. You could compare a collection of texts on a specific topic/discipline (target corpus) with a general corpus (reference corpus) to identify keywords that are relevant to that topic/ discipline.

Keyword analysis reveals words that appear more frequently in the target corpus compared to the reference corpus based on the relative corpus size and some statistical calculations. AntConc's Corpus Manager allows you to manage and select target and reference corpora for analysis. It also offers online wordlist corpora, under the default list, that you can download and use as a reference corpus. These include the BE06 corpus and the AME06 corpus word frequency lists (see Figure 6).

Figure 6

AntConc: Keyword



Status: Deleted default/demo.db

To start keyword analysis:

1. Open Corpus Manager and load your target corpus and reference corpus; use the "Target Corpus" and "Reference Corpus" tabs at the top right of the interface to switch between them. The Target Corpus is the corpus you want to analyze or compare; the Reference Corpus is the corpus that serves as a benchmark for comparison.
2. Go back to the main interface, and click on the "Keyword" tab.
3. Go to "sort by" and select "Likelihood". This option ranks Keywords by the statistical significance of the difference. Based on the corpus size and the results, you may want to focus on the top 20-50 keywords, or you may have a long list of significant results.
4. Change your sort options to reveal different layers of your data. For example, search "students" then flip between Likelihood and Range (Tar) to see if this keyword is both distinctive and consistently distributed across your corpus.

Collocation Analysis

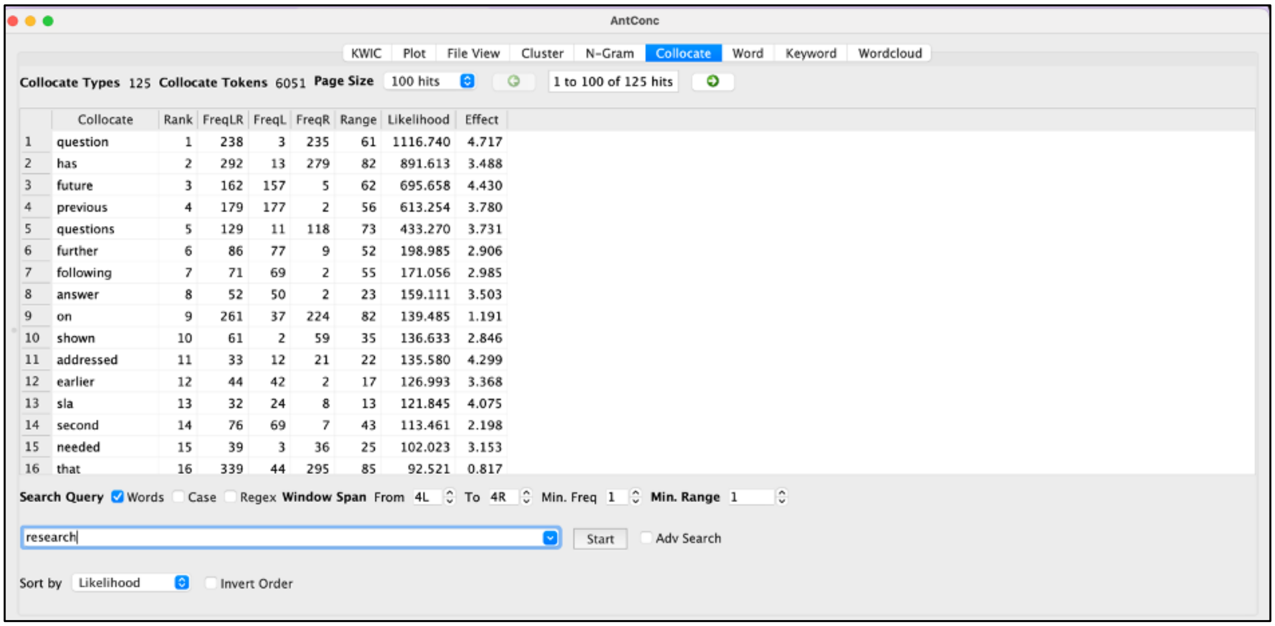
Collocation refers to the tendency of certain words to co-occur more frequently than would be expected by chance. You can look at words that occur directly next to your search word (adjacent collocations) or those occurring within a span (window collocations). As explained by Sinclair et al. (2004), the analysis of collocation involves: node (the target word being studied), collocate (words that appear near the node within a certain distance), span (the window of words examined around the node, e.g., four words before and after), and strength (using statistical calculations to determine whether word pairs co-occur together significantly more frequently than would be expected by chance alone, based on their frequencies and the size of the corpus). Researchers have identified collocations using a collocation-via-concordance approach, where the tool is used for generating the list and the analysis is completed manually without testing for statistical significance, or collocation-via-significance (McEnery & Hardie, 2011).

Here is how to start this analysis:

1. Load your corpus into AntConc
2. Click the Collocate tab
3. Type your node word in the search box. The choice of the word can be driven by the research question, results of frequency or keyword analysis, or informed by literature, technical terminology in a domain, or any other purpose.
4. Set statistical parameters: set window size (e.g., 4L, 4R), minimum frequency (e.g., five occurrences), and sort results by Log-likelihood (i.e., built-in statistical measure used to determine which words are significant collocates)
5. Click ‘start’ to run the analysis.
6. Set the same statistical parameters except for the sorting option: click effect to list the words based on the strength of association (words with a high effect size are strong collocations)
7. Click ‘start’ to run the analysis. When searching for a word’s collocates, a long list of co-occurring words may appear. However, not all of them are meaningful or relevant, so it’s important to filter the results using statistical measures and by reviewing concordance lines in context (Figure 7).
8. Consider those collocates with high Log-likelihood and high effect size
9. Check the concordance lines of the collocates of the given node, and look for patterns

Figure 7

AntConc: Collocate



Collocate Types 125 Collocate Tokens 6051 Page Size 100 hits 1 to 100 of 125 hits

	Collocate	Rank	FreqLR	FreqL	FreqR	Range	Likelihood	Effect
1	question	1	238	3	235	61	1116.740	4.717
2	has	2	292	13	279	82	891.613	3.488
3	future	3	162	157	5	62	695.658	4.430
4	previous	4	179	177	2	56	613.254	3.780
5	questions	5	129	11	118	73	433.270	3.731
6	further	6	86	77	9	52	198.985	2.906
7	following	7	71	69	2	55	171.056	2.985
8	answer	8	52	50	2	23	159.111	3.503
9	on	9	261	37	224	82	139.485	1.191
10	shown	10	61	2	59	35	136.633	2.846
11	addressed	11	33	12	21	22	135.580	4.299
12	earlier	12	44	42	2	17	126.993	3.368
13	sla	13	32	24	8	13	121.845	4.075
14	second	14	76	69	7	43	113.461	2.198
15	needed	15	39	3	36	25	102.023	3.153
16	that	16	339	44	295	85	92.521	0.817

Search Query ☑ Words □ Case □ Regex Window Span From 4L To 4R Min. Freq 1 Min. Range 1

research

Start □ Adv Search

Sort by Likelihood □ Invert Order

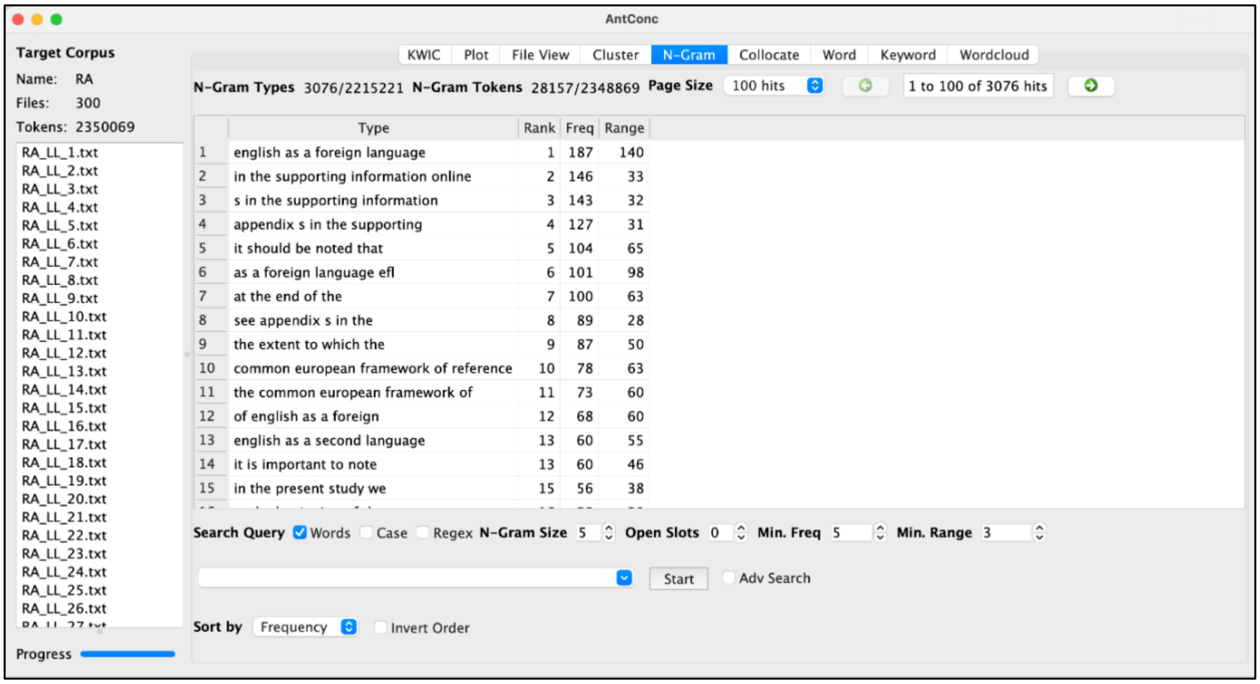
Lexical Bundles

Lexical bundles are continuous sequences of three or more words that occur with high frequency (Biber et al., 1999). You can use the N-Gram tool to extract lexical bundles in your corpus.

Here is how to start this analysis:

1. Load your corpus into AntConc
2. Click the N-Gram tab
3. Set your parameters: decide on the N-gram size (usually 3-6 words, 4 and 5 grams are most common), minimum frequency (e.g., if you are searching for bundles with more than 5 occurrences, set your threshold at 6), and range (distribution of bundles or the number of texts in which the identified bundles occur)
4. Generate the list: a frequency-ranked list of all n-grams meeting your criteria
5. Sort by frequency to see the most common bundles
6. The result will often show a long list of bundles (Figure 8). Copy and paste the results into an Excel file to help organize and filter the data. From the total list of the retrieved lexical bundles, you could exclude those that are incomplete or cross sentence/clausal boundaries. Also, identify and manage overlapping sequences to avoid inflating frequency counts (Chen & Baker, 2010).
7. Analyze the bundles based on their structure, function, or both, and use the concordance function to examine them in context

Figure 8
AntConc: N-Gram



Target Corpus
Name: RA
Files: 300
Tokens: 2350069

RA_LL_1.txt
RA_LL_2.txt
RA_LL_3.txt
RA_LL_4.txt
RA_LL_5.txt
RA_LL_6.txt
RA_LL_7.txt
RA_LL_8.txt
RA_LL_9.txt
RA_LL_10.txt
RA_LL_11.txt
RA_LL_12.txt
RA_LL_13.txt
RA_LL_14.txt
RA_LL_15.txt
RA_LL_16.txt
RA_LL_17.txt
RA_LL_18.txt
RA_LL_19.txt
RA_LL_20.txt
RA_LL_21.txt
RA_LL_22.txt
RA_LL_23.txt
RA_LL_24.txt
RA_LL_25.txt
RA_LL_26.txt

N-Gram Types 3076/2215221 **N-Gram Tokens** 28157/2348869 **Page Size** 100 hits 1 to 100 of 3076 hits

	Type	Rank	Freq	Range
1	english as a foreign language	1	187	140
2	in the supporting information online	2	146	33
3	s in the supporting information	3	143	32
4	appendix s in the supporting	4	127	31
5	it should be noted that	5	104	65
6	as a foreign language efl	6	101	98
7	at the end of the	7	100	63
8	see appendix s in the	8	89	28
9	the extent to which the	9	87	50
10	common european framework of reference	10	78	63
11	the common european framework of	11	73	60
12	of english as a foreign	12	68	60
13	english as a second language	13	60	55
14	it is important to note	13	60	46
15	in the present study we	15	56	38

Search Query ☑ Words □ Case □ Regex **N-Gram Size** 5 **Open Slots** 0 **Min. Freq** 5 **Min. Range** 3

□ Adv Search

Sort by Frequency □ Invert Order

POS-tagged Corpus Analysis

After preparing your POS-tagged corpus files, you can load them into AntConc for part-of-speech analysis. This section shows how to set up AntConc to work with POS tags annotated via the Stanford Parser.

Setting up POS-tagged corpus in AntConc

1. Open AntConc and navigate to File → Open Corpus Manager
2. Create a new corpus by clicking “Corpus Source” → “Raw File(s)”
3. Name your corpus (e.g., RA_POS_Tagged)
4. Add your POS-tagged files from your corpus folder using “Add File(s)” or “Add Directory”
5. In Basic Settings, click the arrow next to “Indexer, Encoding, Token Definition, Row Processor”
6. Change the indexer from “simple_word_indexer” to “simple_word_pos_headword_indexer”
7. Click “Create” to establish the corpus with POS processing enabled and return to main window
8. Go to Settings → Global Settings → Tags → configure Display Type options:
   - **Type:** Shows only the word (e.g., student)
   - **Type+POS:** Shows word with POS tag (e.g., student_NN)
   - **Type+Headword:** Shows word with its base form (e.g., running_run)
   - **Type+POS+Headword:** Shows word, POS tag, and base form (e.g., running_VBG_run)
   - **POS:** Shows only the POS tag (e.g., NN)
   - **Headword:** Shows only the base form (e.g., run)

- **Headword+POS:** Shows base form with POS tag (e.g., run_VBG)

For most POS analysis tasks, select "Type+POS"

9. Click "Set for all tools" to apply the configuration across AntConc
10. Click "Apply" to save the settings

Using POS Tags for Analysis

With POS-tagged corpora loaded, you can perform targeted analyses using specific search patterns:

Part-of-Speech Specific Word Lists: Use the Word tool with POS search patterns to generate frequency lists for specific grammatical categories. In the following categories, an * is used to represent any number of characters in the word, and the Penn Treebank tagset was consulted to identify target tags. For Example, in the first pattern below "*_" allows for any word before the "_", "NN*" allows for all of the noun tags to be included, as they all *start* with NN, but some of them end with other characters.

- Search for *_NN* to display all nouns
- **Search for * _JJ* to display all adjectives** (Figure 9)
- Search for *_VB* to display all verbs
- Search for *_IN* to display all prepositions

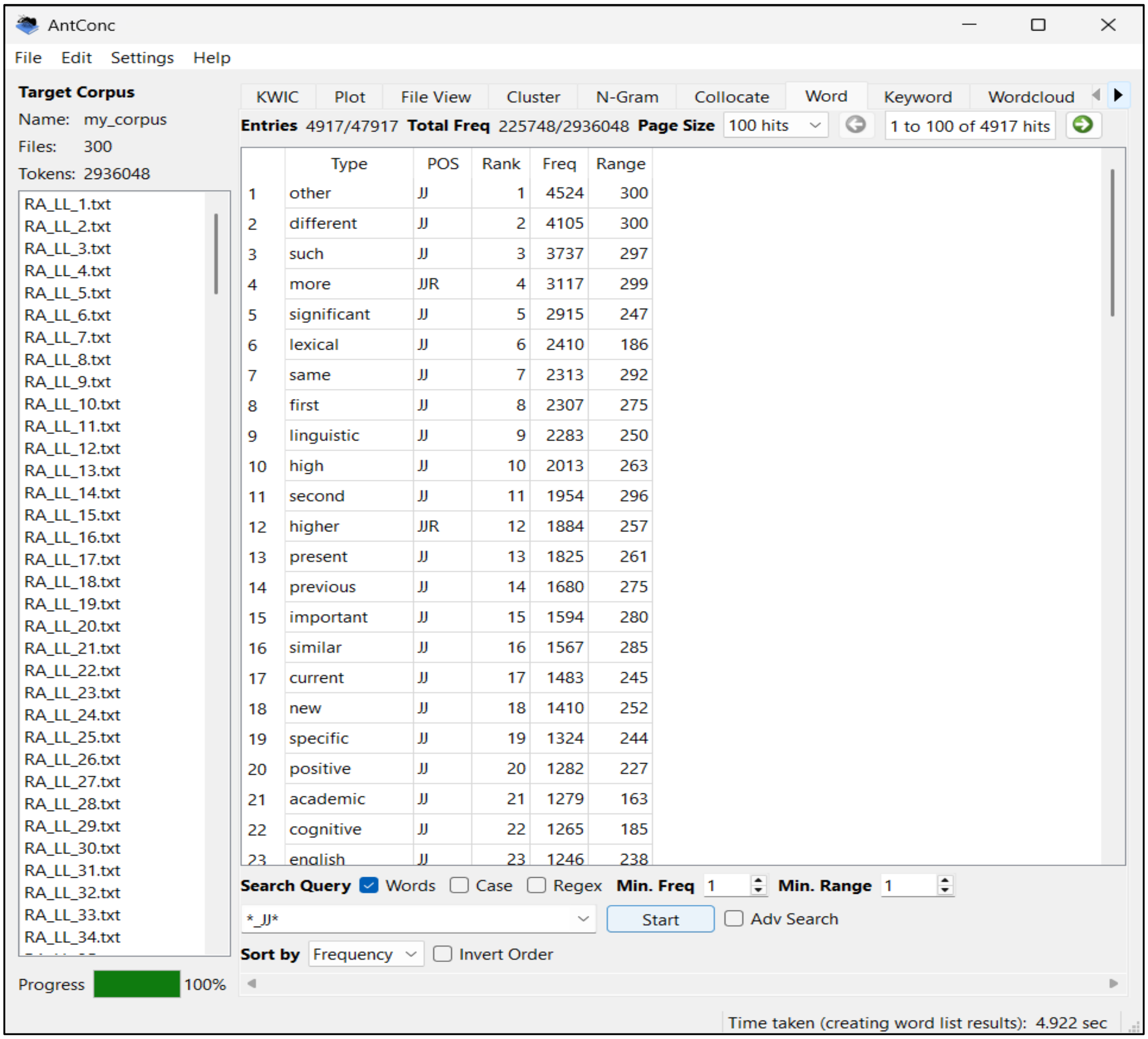
Collocate Analysis with POS Filtering: Use the Collocate tool to search for specific POS patterns around target words:

- Use *_JJ* student* to find all adjectives within the designated window of "student" (Figure 10)

Complex Pattern Searches: Use the KWIC tool to combine multiple POS tags and identify specific grammatical structures:
*_JJ * *_NN * finds all adjective-noun sequences (Figure 11)

Figure 9

*AntConc Adjective Frequency List (*_JJ*)*



Target Corpus

Name: my_corpus

Files: 300

Tokens: 2936048

RA_LL_1.txt
RA_LL_2.txt
RA_LL_3.txt
RA_LL_4.txt
RA_LL_5.txt
RA_LL_6.txt
RA_LL_7.txt
RA_LL_8.txt
RA_LL_9.txt
RA_LL_10.txt
RA_LL_11.txt
RA_LL_12.txt
RA_LL_13.txt
RA_LL_14.txt
RA_LL_15.txt
RA_LL_16.txt
RA_LL_17.txt
RA_LL_18.txt
RA_LL_19.txt
RA_LL_20.txt
RA_LL_21.txt
RA_LL_22.txt
RA_LL_23.txt
RA_LL_24.txt
RA_LL_25.txt
RA_LL_26.txt
RA_LL_27.txt
RA_LL_28.txt
RA_LL_29.txt
RA_LL_30.txt
RA_LL_31.txt
RA_LL_32.txt
RA_LL_33.txt
RA_LL_34.txt

Progress 100%

Entries 4917/47917 **Total Freq** 225748/2936048 **Page Size** 100 hits

1 to 100 of 4917 hits

	Type	POS	Rank	Freq	Range
1	other	JJ	1	4524	300
2	different	JJ	2	4105	300
3	such	JJ	3	3737	297
4	more	JJR	4	3117	299
5	significant	JJ	5	2915	247
6	lexical	JJ	6	2410	186
7	same	JJ	7	2313	292
8	first	JJ	8	2307	275
9	linguistic	JJ	9	2283	250
10	high	JJ	10	2013	263
11	second	JJ	11	1954	296
12	higher	JJR	12	1884	257
13	present	JJ	13	1825	261
14	previous	JJ	14	1680	275
15	important	JJ	15	1594	280
16	similar	JJ	16	1567	285
17	current	JJ	17	1483	245
18	new	JJ	18	1410	252
19	specific	JJ	19	1324	244
20	positive	JJ	20	1282	227
21	academic	JJ	21	1279	163
22	cognitive	JJ	22	1265	185
23	english	JJ	23	1246	238

Search Query ☑ Words □ Case □ Regex **Min. Freq** 1 **Min. Range** 1

_JJ

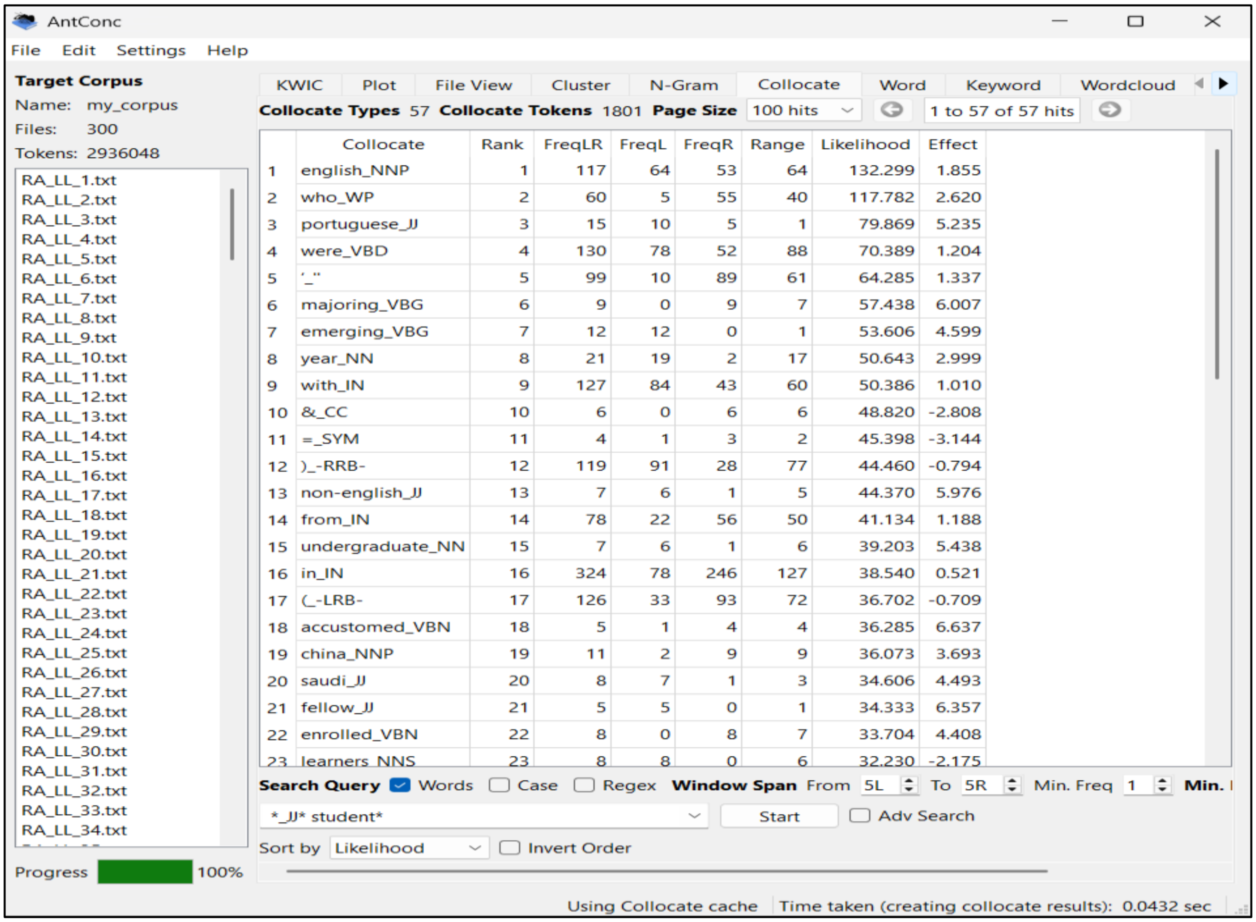
□ Adv Search

Sort by Frequency □ Invert Order

Time taken (creating word list results): 4.922 sec

Figure 10

*AntConc Collocate Results for (*_JJ* student*)*


AntConc
File Edit Settings Help
KWIC Plot File View Cluster N-Gram Collocate Word Keyword Wordcloud
Start

Target Corpus
Name: my_corpus
Files: 300
Tokens: 2936048

RA_LL_1.txt
RA_LL_2.txt
RA_LL_3.txt
RA_LL_4.txt
RA_LL_5.txt
RA_LL_6.txt
RA_LL_7.txt
RA_LL_8.txt
RA_LL_9.txt
RA_LL_10.txt
RA_LL_11.txt
RA_LL_12.txt
RA_LL_13.txt
RA_LL_14.txt
RA_LL_15.txt
RA_LL_16.txt
RA_LL_17.txt
RA_LL_18.txt
RA_LL_19.txt
RA_LL_20.txt
RA_LL_21.txt
RA_LL_22.txt
RA_LL_23.txt
RA_LL_24.txt
RA_LL_25.txt
RA_LL_26.txt
RA_LL_27.txt
RA_LL_28.txt
RA_LL_29.txt
RA_LL_30.txt
RA_LL_31.txt
RA_LL_32.txt
RA_LL_33.txt
RA_LL_34.txt

Progress 100%

Collocate Types 57 **Collocate Tokens** 1801 **Page Size** 100 hits 1 to 57 of 57 hits

	Collocate	Rank	FreqLR	FreqL	FreqR	Range	Likelihood	Effect
1	english_NNP	1	117	64	53	64	132.299	1.855
2	who_WP	2	60	5	55	40	117.782	2.620
3	portuguese_JJ	3	15	10	5	1	79.869	5.235
4	were_VBD	4	130	78	52	88	70.389	1.204
5	‘_"	5	99	10	89	61	64.285	1.337
6	majoring_VBG	6	9	0	9	7	57.438	6.007
7	emerging_VBG	7	12	12	0	1	53.606	4.599
8	year_NN	8	21	19	2	17	50.643	2.999
9	with_IN	9	127	84	43	60	50.386	1.010
10	&_CC	10	6	0	6	6	48.820	-2.808
11	=_SYM	11	4	1	3	2	45.398	-3.144
12	)_-RRB-	12	119	91	28	77	44.460	-0.794
13	non-english_JJ	13	7	6	1	5	44.370	5.976
14	from_IN	14	78	22	56	50	41.134	1.188
15	undergraduate_NN	15	7	6	1	6	39.203	5.438
16	in_IN	16	324	78	246	127	38.540	0.521
17	(_-LRB-	17	126	33	93	72	36.702	-0.709
18	accustomed_VBN	18	5	1	4	4	36.285	6.637
19	china_NNP	19	11	2	9	9	36.073	3.693
20	saudi_JJ	20	8	7	1	3	34.606	4.493
21	fellow_JJ	21	5	5	0	1	34.333	6.357
22	enrolled_VBN	22	8	0	8	7	33.704	4.408
23	learners_NNS	23	8	8	0	6	32.230	-2.175

Search Query ☑ Words ☐ Case ☐ Regex **Window Span** From 5L To 5R Min. Freq 1 **Min.**

_JJ student*

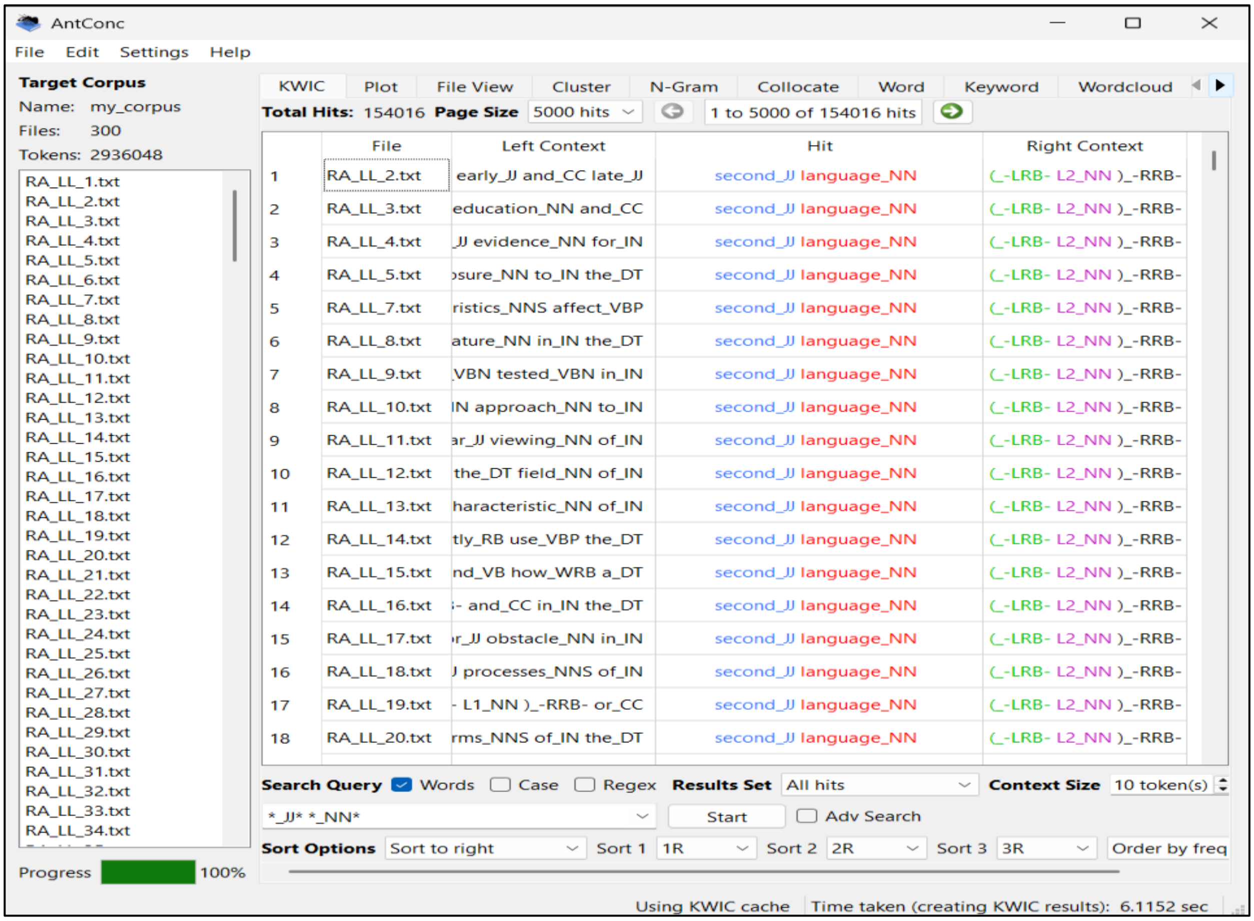
☐ Adv Search

Sort by Likelihood ☐ Invert Order

Using Collocate cache | Time taken (creating collocate results): 0.0432 sec

Figure 11

*AntConc KWIC Results for (*_JJ * *_NN *) Pattern Search*


AntConc
File Edit Settings Help
KWIC Plot File View Cluster N-Gram Collocate Word Keyword Wordcloud

Target Corpus
Name: my_corpus
Files: 300
Tokens: 2936048

RA_LL_1.txt
RA_LL_2.txt
RA_LL_3.txt
RA_LL_4.txt
RA_LL_5.txt
RA_LL_6.txt
RA_LL_7.txt
RA_LL_8.txt
RA_LL_9.txt
RA_LL_10.txt
RA_LL_11.txt
RA_LL_12.txt
RA_LL_13.txt
RA_LL_14.txt
RA_LL_15.txt
RA_LL_16.txt
RA_LL_17.txt
RA_LL_18.txt
RA_LL_19.txt
RA_LL_20.txt
RA_LL_21.txt
RA_LL_22.txt
RA_LL_23.txt
RA_LL_24.txt
RA_LL_25.txt
RA_LL_26.txt
RA_LL_27.txt
RA_LL_28.txt
RA_LL_29.txt
RA_LL_30.txt
RA_LL_31.txt
RA_LL_32.txt
RA_LL_33.txt
RA_LL_34.txt

Progress 100%

Total Hits: 154016 **Page Size** 5000 hits 1 to 5000 of 154016 hits

	File	Left Context	Hit	Right Context
1	RA_LL_2.txt	early_JJ and_CC late_JJ	second_JJ language_NN	(_-LRB- L2_NN)_-RRB-
2	RA_LL_3.txt	education_NN and_CC	second_JJ language_NN	(_-LRB- L2_NN)_-RRB-
3	RA_LL_4.txt	JJ evidence_NN for_IN	second_JJ language_NN	(_-LRB- L2_NN)_-RRB-
4	RA_LL_5.txt	osure_NN to_IN the_DT	second_JJ language_NN	(_-LRB- L2_NN)_-RRB-
5	RA_LL_7.txt	ristics_NNS affect_VBP	second_JJ language_NN	(_-LRB- L2_NN)_-RRB-
6	RA_LL_8.txt	ature_NN in_IN the_DT	second_JJ language_NN	(_-LRB- L2_NN)_-RRB-
7	RA_LL_9.txt	VBN tested_VBN in_IN	second_JJ language_NN	(_-LRB- L2_NN)_-RRB-
8	RA_LL_10.txt	IN approach_NN to_IN	second_JJ language_NN	(_-LRB- L2_NN)_-RRB-
9	RA_LL_11.txt	ar_JJ viewing_NN of_IN	second_JJ language_NN	(_-LRB- L2_NN)_-RRB-
10	RA_LL_12.txt	the_DT field_NN of_IN	second_JJ language_NN	(_-LRB- L2_NN)_-RRB-
11	RA_LL_13.txt	haracteristic_NN of_IN	second_JJ language_NN	(_-LRB- L2_NN)_-RRB-
12	RA_LL_14.txt	tly_RB use_VBP the_DT	second_JJ language_NN	(_-LRB- L2_NN)_-RRB-
13	RA_LL_15.txt	nd_VB how_WRB a_DT	second_JJ language_NN	(_-LRB- L2_NN)_-RRB-
14	RA_LL_16.txt	- and_CC in_IN the_DT	second_JJ language_NN	(_-LRB- L2_NN)_-RRB-
15	RA_LL_17.txt	r_JJ obstacle_NN in_IN	second_JJ language_NN	(_-LRB- L2_NN)_-RRB-
16	RA_LL_18.txt	J processes_NNS of_IN	second_JJ language_NN	(_-LRB- L2_NN)_-RRB-
17	RA_LL_19.txt	- L1_NN)_-RRB- or_CC	second_JJ language_NN	(_-LRB- L2_NN)_-RRB-
18	RA_LL_20.txt	rms_NNS of_IN the_DT	second_JJ language_NN	(_-LRB- L2_NN)_-RRB-

Search Query ☑ Words □ Case □ Regex **Results Set** All hits **Context Size** 10 token(s)

_JJ *_NN* Start □ Adv Search

Sort Options Sort to right Sort 1 1R Sort 2 2R Sort 3 3R Order by freq

Using KWIC cache Time taken (creating KWIC results): 6.1152 sec

Phrase-frame Analysis with KfNgram

Another interesting kind of multiword sequences is called phrase-frames. Phrase-frames are discontinuous sequences with a variable slot that can be filled by semantically coherent variants (Fletcher, 2012). We recommend using kfNgram, a free downloadable Windows software, which you can download from http://www.kwicfinder.com/kfNgram/kfNgramStopwords.exe. More information about kfNgram and its functions can be found here:
https://www.kwicfinder.com/kfNgram/kfNgramHelp.html.

First, you need to combine your text files into a single file in order to extract all word n-grams and phrase frames from one unified source. The following will show you how to do this on both Windows and Mac.

To combine files on Windows:

1. Click on the folder containing the files you want to combine.

2. Hold down Shift and right-click in the folder.
3. Click Open PowerShell window here
4. Type Get-Content *.txt | Set-Content new file.txt and press Enter to combine (Figure 12)
5. Check the folder. You will find a new file called "newfile.txt" created there with the combined contents.

To combine files on Mac:

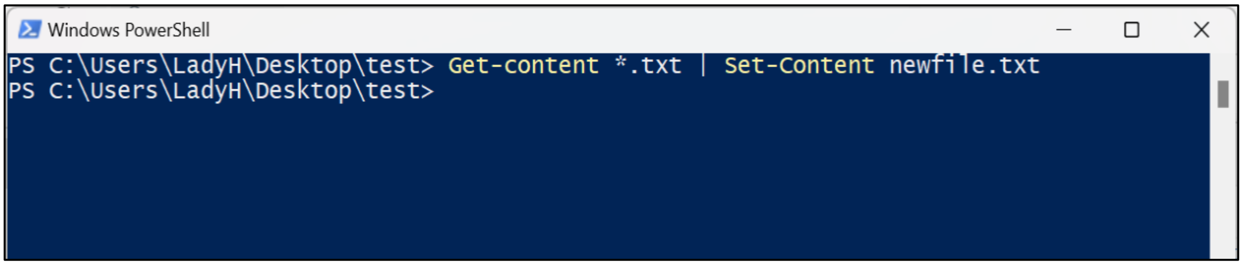
1. Click on the folder containing the files you want to combine.
2. Hold down Shift and right-click in the folder.
3. Click New Terminal at Folder
4. Type cat *.txt > combined.txt (Figure 13)
5. Check the folder. You will find a new file called "combined.txt" created there with the combined contents.

To extract phrase-frames:

1. Open kfNgram
2. Set your parameters: decide on the N-gram size (usually 4-6 words), floor 1, not case-sensitive, replace .,-' with space, frequency sort, retain numerals, separate
3. Add the source file (the combined text file)
4. Go to Tools, and select Get Wordgrams (Figure 14)
5. Open a new kfNgram window, use the same settings: floor 1, not case-sensitive, replace .,-' with space, frequency sort, retain numerals, separate
6. Add the source file (the Wordgram file generated by kfNgram in the previous step), if it does not show you the file (No items match your research), check the file type and change it from Alpha n-Gram files to frequency nGram files
7. Go to tools and select Get Phrase-Frames from Wordgram file (Figure 14)
8. From the output text file, copy only the phrase frames that match your minimum frequency threshold (e.g., those which have 3 or 5 occurrences) and paste them into an Excel file. Note that there are two numbers next to each phrase frame. The first number indicates how many times the phrase frame appears in the corpus. This is the one you need to check for minimum frequency. The second number tells you how many types of variants fill this phrase frame. Then, you have the number of occurrences for each variant.
9. In the Excel file, go through the list and clean the frames and variants. When cleaning them, remove those that contain repeated words, incomplete patterns, or nonsensical word combinations. Focus on frames that are grammatically sound, meaningful, and meet the predetermined frequency and range threshold. Also, make sure to identify and manage overlapping sequences. Start by cleaning the frames, then clean the variants, and create a separate sheet for each round of cleaning. As you clean the phrases and the variants, use AntConc along with all the individual text files (the ones you combined to extract the phrase frames) to check the context and the range (number of files in which the phrase occurs). As you clean, update the frames, variants, and their frequencies.
10. Analyze the results based on their structure, function, or both, and use the concordance function to examine phrase frames in context

Figure 12

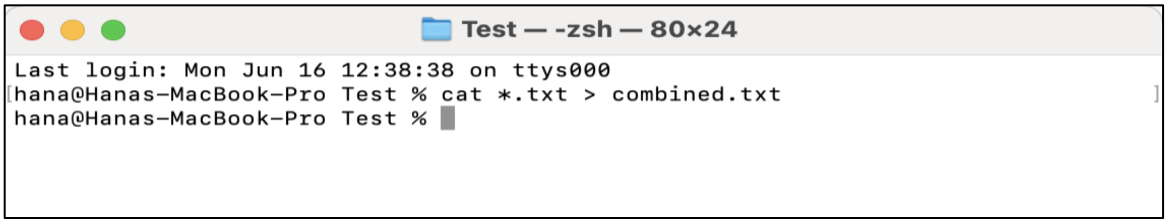
Windows PowerShell: Combine Files



```
PS C:\Users\LadyH\Desktop\test> Get-content *.txt | Set-Content newfile.txt
PS C:\Users\LadyH\Desktop\test>
```

Figure 13

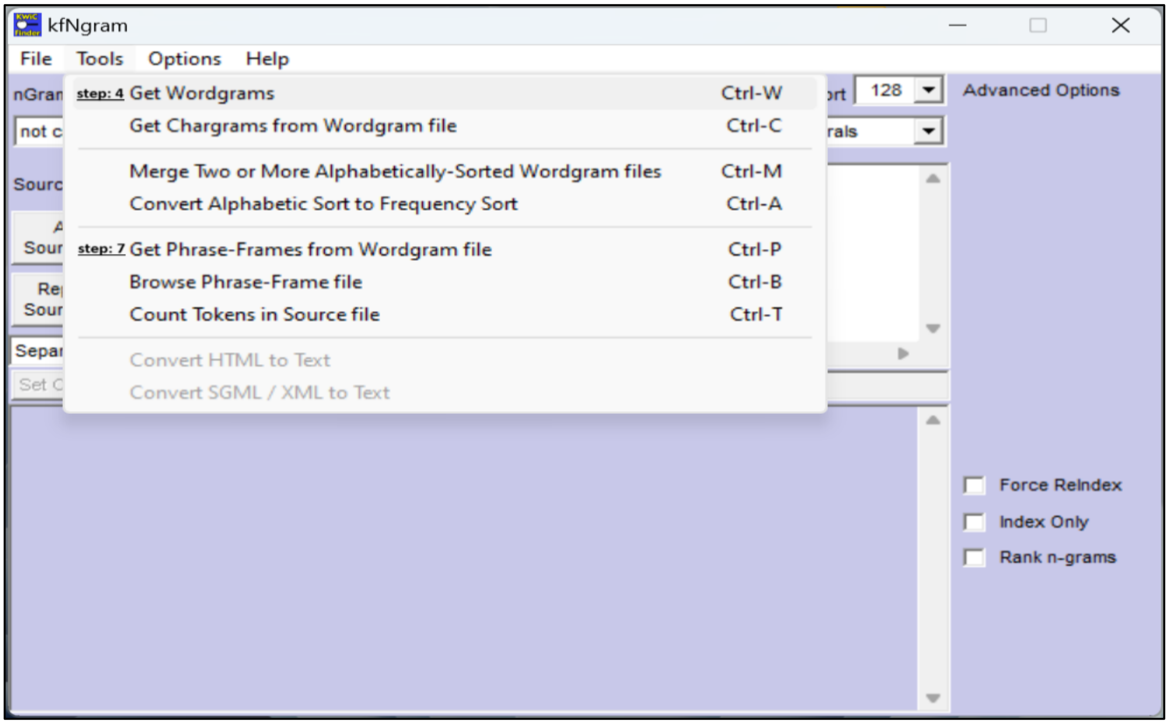
Mac Terminal: Combine Files



```
Last login: Mon Jun 16 12:38:38 on ttys000
hana@Hanas-MacBook-Pro Test % cat *.txt > combined.txt
hana@Hanas-MacBook-Pro Test %
```

Figure 14

KfNgram: Phrase Frames Extraction_ Steps 4 & 7


kfNgram
File
Tools
Options
Help
step: 4 Get Wordgrams Ctrl-W
Get Chargrams from Wordgram file Ctrl-C
Merge Two or More Alphabetically-Sorted Wordgram files Ctrl-M
Convert Alphabetic Sort to Frequency Sort Ctrl-A
step: 7 Get Phrase-Frames from Wordgram file Ctrl-P
Browse Phrase-Frame file Ctrl-B
Count Tokens in Source file Ctrl-T
Convert HTML to Text
Convert SGML / XML to Text
128
Advanced Options
Force ReIndex
Index Only
Rank n-grams

Conclusion

The step-by-step instructions provided in this guide for selecting existing corpora or building custom ones and then analyzing them linguistically enable language teachers to enhance their pedagogical practice in multiple ways (O'Keeffe & McCarthy, 2022). In materials and curriculum development, teachers can use corpus data to develop vocabulary and grammar content based on frequency patterns and authentic usage rather than intuition-based approaches. Teachers can apply the same techniques to collect and analyze their students' output, e.g., writing samples, to identify common errors and discover what grammatical structures students have mastered and which areas need attention. This analysis enables them to design lessons that address actual needs rather than relying on traditional sequencing. In classroom applications, teachers can implement Data-driven learning (DDL) tasks where students use concordancing software such as AntConc to learn about language use by analyzing teacher-prepared concordances or by independently exploring corpora. The typical DDL tasks present concordances with guiding questions to recognize patterns and draw conclusions. Teachers can design tasks to teach specific linguistic features or raise awareness of how certain words or phrases function in context. In test development, teachers can use corpus data to write test items appropriate for students' proficiency levels and based on real language use. For example, by analyzing learner corpora, teachers can determine which vocabulary, grammar, and language patterns are typical of each proficiency level and select appropriate ones for their tests. These applications demonstrate that corpus analysis can inform and develop various aspects of language teaching.

In this guide, we have shown that building, cleaning, and analyzing corpora is systematic and manageable. The step-by-step instructions we provided make corpus work accessible to language teachers and novice researchers. The instructions in this guide can be used as a starting point, but expect to adapt them based on any teaching and/or research needs.

We encourage new analysts to start with small projects to build confidence before working with larger corpora. Corpus linguistics is an iterative process. The first corpus teaches lessons that improve the second project. Each analysis helps build understanding for what works best for specific research questions. Any challenges or problems that arise should not be seen as failures, but as a part of the learning process. Over time, it becomes clear that some cleaning decisions need changes, that certain analytical approaches work better than others, and that corpus work often leads to unexpected findings.

Acknowledgments

We would like to thank Dr. J. Elliott Casal for his guidance and mentorship in learning how to do corpus-based research.

Authors

Ghadi Matouq is a Doctoral Candidate in Applied Linguistics at the University of Memphis and Lecturer in the English Department at Taif University. She holds a Master's in Applied Linguistics and Teaching English in International Contexts (TEIC) certificate from Texas Tech University. Her research interests include corpus linguistics, professional discourse, and computer-assisted language learning.

Hana Alqabba is a Doctoral Candidate in Applied Linguistics at the University of Memphis and Lecturer in the Department of English Language and Translation at Qassim University. Her research interests include corpus linguistics, genre analysis, and academic writing, with a focus on discipline- and paradigm-specific writing practices and literacies.

References

Anthony, L. (2023). AntConc (Version 4.2.4) [Computer Software]. Tokyo, Japan: Waseda University. https://www.laurenceanthony.net/software/AntConc

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Longman.

Chen, Y. H., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning and Technology, 14*, 30–49.

Egbert, J., Larsson, T., & Biber, D. (2020). *Doing linguistics with a corpus: Methodological considerations for the everyday user*. Cambridge University Press.

Fletcher, W. H. (2012). *KfNgram*. Annapolis MD: USA.

Foster, M. E., & Oberlander, J. (2007). Corpus-based generation of head and eyebrow motion for an embodied conversational agent. *Language Resources and Evaluation*, *41*, 305–323.

Hyland, K. (1999). Academic attribution: Citation and the construction of disciplinary knowledge. *Applied linguistics*, *20*(3), 341–367.

Hyland, K. (2008). 'Small bits of textual material': A discourse analysis of Swales' writing. *English for Specific Purposes*, *27*(2), 143-160.

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., & Suchomel, V. (2013, July). The TenTen corpus family. In 7th international corpus linguistics conference CL (Vol. 2013, pp. 125-127).

Lu, X. (2014). *Computational methods for corpus annotation and analysis*. Springer.

McCarthy, M., & O'Keeffe, A. (2010). Historical perspective: What are corpora and how have they evolved? In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 3–13). Routledge.

McEnery, T., & Hardie, A. (2011). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.

Newman, J., & Cox, C. (2020). Corpus annotation. In M. Lefer, M. Paquot, & S. T. Gries (Eds.), *A practical handbook of corpus linguistics* (pp. 25-48). Springer.

O'Keeffe, A., & McCarthy, M.J. (Eds.). (2022). *The Routledge handbook of corpus linguistics* (2nd ed.). Routledge. https://doi.org/10.4324/9780367076399

Reppen, R., & Simpson, R. (2002). Corpus linguistics. In N. Schmitt (Ed.), *An Introduction to applied linguistics* (pp.89–106). Arnold.

Reppen, R. (2022). Building a corpus: What are key considerations?. In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 13-20). Routledge. https://doi.org/10.4324/9780367076399

Sinclair, J. (2004). Corpus and text: Basic principles. In M. Wynne (Ed.), *Developing linguistic corpora: A guide to good practice*. Oxford. https://doi.org/20.500.14106/dlc

Sinclair, J., Jones, S., & Daley, R. (2004). *English collocation studies: The OSTI report*. Bloomsbury Publishing.

Timmis, I. (2015). *Corpus linguistics for ELT: Research and practice* (1st ed.). Routledge. https://doi.org/10.4324/9781315715537

Thompson, P. (2004). Spoken language corpora. In M. Wynne (Ed.), *Developing linguistic corpora: A guide to good practice*. Oxford. https://doi.org/20.500.14106/dlc

Appendix

Importance of Text Preparation and Cleaning

Converting raw textual data into a clean, standardized, and analysis-ready format is a crucial task before analyzing spoken or written language (Reppen, 2022). The following problems demonstrate why proper text preparation and cleaning are essential:

- **File organization failures:** As a first attempt you may want to put multiple texts in a large file, but this limits your analysis options because you cannot analyze texts individually (e.g., students' essays), group them by different criteria (such as gender, proficiency level, or time period), or combine them as needed for different research questions. This forces you into the time-consuming process of later dividing large files and resaving them with new names or rebuilding the corpus.
- **Poor naming conventions:** File names that do not clearly reflect the content make finding and organizing texts extremely difficult, while longer names can create compatibility issues across analytical software and backup systems.
- **PDF conversion issues:** Converting PDF files to text format can disrupt paragraph sequencing and scramble words or characters. Tools like Wondershare (https://shorturl.at/3IBqp) allow PDF editing and conversion to TXT format but may sometimes disrupt paragraph sequencing. Voyant Tools (https://shorturl.at/SSXZM) preserves paragraph sequencing during conversion, and TextFixer (https://shorturl.at/pe8jE) along with Capitalize My Title (https://shorturl.at/saRm8) help remove unwanted line breaks from texts. Generally, you need to compare with the original PDF file to check for any conversion errors or formatting issues.
- **Data loss risks:** Without backup copies of texts in multiple secure locations, you risk losing files from computer crashes, fires, or theft. These things do happen.
- **Missing metadata:** Lost headers or metadata means you lose valuable contextual information for future analysis, and if headers are not properly enclosed with angle brackets and separators, this information could be accidentally included in text analysis, which could inflate frequency counts.
- **Format incompatibility:** Files saved in incompatible formats create significant problems during analysis and require additional processing time, as plain text or UTF8 formats work best with corpus analysis tools.

- **Non-standard language handling:** When working with written texts that contain non-standard language, such as learner language, novice writing, or children's writing, you may want to create two versions of each text. An original version preserves all spelling and grammar structures because they may be valuable for certain types of analysis, and a clean version with standardized spelling for other analyses.
- **Poor recording quality:** When building spoken corpora, failing to ensure high-quality recordings using digital devices such as phones and tablets results in unclear sound that does not transfer easily to computers and creates major transcription difficulties.
- **Inconsistent transcription decisions:** You must make important decisions about how to handle reduced speech forms, unclear audio segments, overlapping speech, and prosodic elements consistently throughout the corpus.