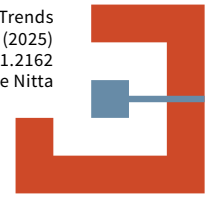
Trends

JALTCALL Trends
ISSN 2982-3749, 1, 2162 (2025)
https://doi.org/10.29140/jct.v1n1.2162



Castledown

An introduction to corpus technology in the age of AI

Sadie Nitta
University of Electro-Communications, Japan
sadienitta@uec.ac.jp

Corpus technology has been noted as a facilitative tool in data-driven learning (DDL) due to its advanced search capabilities and access to vast linguistic databases. Despite its potential, many teachers shy away from direct corpus use, which involves teacher- and learner-corpus interaction. Based on teacher and learner feedback, this hesitation is due to the overwhelming amount of corpus data and unfamiliarity with corpus query functions and techniques. To address these questions, this practice-based article aims to bridge the gap between teachers and the evolving landscape of corpus technology in the age of AI. First, it provides a brief introduction to the term corpus and its associated technology. Second, it highlights the common types of corpora available for teachers as an open resource. Next, it addresses corpus basics using Anthony's AntConc 4.3.1 software. Then, it suggests three simple ways to incorporate corpus technology in the classroom. Finally, the article summarizes a corpus-informed lesson following the Illustration-Induction-Interaction (I-I-I) model.

Keywords: corpus, data-driven learning, open resources, AntConc

Introduction

What is a *corpus?* According to Merriam-Webster.com, the term corpus derives from Latin, meaning "body" or "collection" ("corpus," 2024). The origin is fitting, given its resemblance to the term *corpse*. To modern linguists, in the age of AI, a corpus is formally known as a large, principled collection of authentic, naturally occurring spoken or written language in machine-readable text (Biber et al., 2007). Additionally, corpus technology has been noted as a powerful, facilitative tool in data-driven learning (DDL) (Johns, 1991), valued for its advanced search capabilities and access to vast linguistic databases. Despite its potential, many teachers shy away from direct corpus use, often due to the overwhelming amount of corpus data and unfamiliarity with corpus query functions and techniques (Boulton, 2010). For this reason, *corpora* should be recognized as a special type of informant, not a tutor or surrogate teacher (Johns, 1991). With these challenges in mind, this practice-based article aims to bridge the gap between teachers and the evolving landscape of corpus technology

by summarizing accessible corpus types, basics, uses, and applications for the language classroom.

Corpus types

General or reference corpora

A general or reference corpus is the most extensive collection of natural texts from the target language, such as North American English or British English (Friginal, 2018). The database is not restricted to one register or domain, capturing a language's birth, life, and sometimes death. A commonly used general or reference corpus is:

The Corpus of Contemporary American English (COCA). Created by Davies (2008-present), COCA (https://www.english-corpora.org/coca/) spans over one billion words – 1,001,610,938, to be exact – and is divided into eight genres: blogs (125 million words [125,496,215]); web pages (130 million words [129,899,426]); TV/movie subtitles (128 million words [128,013,334]); spoken (127 million words [127,396,916]); fiction (119 million words [119,505,292]); popular magazines (127 million words [127,352,014]); newspapers (123 million words [122,959,393]); and academic journals (121 million words [120,988,348]).

Specialized corpora

A specialized corpus is a collection of natural texts from the target language in specific subject areas, such as Academic Spoken English or Business English (Friginal, 2018). The database is limited to one register or domain, allowing in-depth analysis. A commonly used specialized corpus is:

Michigan Corpus of Academic Spoken English (MICASE). Created by the English Language Institute (ELI) at the University of Michigan (1997-2002), MICASE (https://quod.lib.umich.edu/m/micase/) spans over one million words – 1,695,540, to be exact – and is divided into 16 speech event types: advising (58,817 words); colloquia (151,639 words); discussion sections (74,904 words); dissertation defenses (56,837 words); interviews (13,015 words); labs (73,815 words); large lectures (257,311 words); small lectures (320,893 words); meetings (70,038 words); office hours (120,629 words); seminars (151,071 words); study groups (129,725 words); student presentations (143,369 words); service encounters (24,691 words); tours (21,768 words); and tutorials (27,014 words).

Parallel or multilingual corpora

A parallel or multilingual corpus is a collection of natural texts from two or more target languages, such as English and Japanese (Friginal, 2018). The database consists of translations between two or more monolingual corpora, facilitating comparative analysis across languages. A commonly used parallel or multilingual corpus is:

WebParaNews. Created by Anthony and Chujo (n.d.), WebParaNews (https://www.antlabsolutions.com/webparanews/about.html) is a bilingual English and Japanese news corpus.

Learner corpora

A learner corpus is a collection of natural texts produced by language learners, such as Japanese university students studying English (Timmis, 2015). The database showcases the growth and challenges learners face during the acquisition process. A commonly used learner corpus is:

The Michigan Corpus of Upper-level Student Papers (MICUSP). Created by the English Language Institute (ELI) at the University of Michigan (2009), MICUSP (https://elicorpora.info/main) spans over two million words – 2,594,004 words, to be exact – and is divided into 16 disciplines: biology (174,112 words); civil and environmental engineering (98,213 words); economics (77,668 words); education (149,099 words); English (265,396 words); history and classical studies (180,523 words); industrial and operations engineering (124,527 words); linguistics (154,160 words); mechanical engineering (122,176 words); natural resources and environment (172,071 words); nursing (157,785 words); philosophy (126,290 words); physics (44,866 words); political science (208,223 words); psychology (322,808 words); and sociology (214,087 words). Additionally, MICUSP is divided into seven paper types: argumentative essay (22%), creative writing (1%), critique (7%), proposal (6%), report (44%), research paper (17%), and response paper (3%).

Corpus basics

Using Anthony's (2024a) AntConc 4.3.1 software is an approachable way to learn and introduce corpus query functions and techniques. AntConc is a freeware corpus analysis toolkit for concordancing and text analysis. By downloading AntConc 4.3.1, users have access to ten investigative tools (see Figure 1).

TOOL	PURPOSE
1. KWIC (Key-Word-In-Context)	Displays words and phrases in their surrounding context.
2. Plot	Visualizes word frequency in a barcode format.
3. File	Manages corpus files.
4. Cluster	Identifies multiword expressions based on the target word.
5. N-Gram	Generates a list of all N-sized clusters in the corpus.
6. Collocate	Identifies co-occurring words based on the target word.
7. Word List	Generates a list of all words in the corpus.
8. Keyword List	Highlights words that appear unusually frequently compared to a reference corpus.
9. Wordcloud	Creates a visual representation of the most frequent words in the corpus.
10. Chat AI	Provides automated assistance for AntConc tools.

Figure 1. AntConc tools

If learners can access computers during class, the teacher can use printouts of Figure 2 to introduce the corpus technology and guide learners through the generalization process. This lesson extensively uses computers for analysis, incorporating both automatic and interactive techniques (Biber, 1996).

NAME: _____

CLASS: _____

AntConc Basics for EFL Students

❖ Introduction:

A corpus is a database of searchable language that can help you understand English grammar and word patterns. To begin, you must download data (.txt) and a program (such as AntConc 4.3.1) designed to help you search the data.

❖ Objective:

By the end of today's lesson, students will be able to use basic search features in AntConc.

❖ Setting Up:

1. First, create a master folder titled "Corpora."
2. Within the master folder, create two subfolders titled "Corpus Files" (for the data) and "Programs" (for the program).
3. Next, download the corpus files from [ ].
4. Then, download the appropriate version of Laurence Anthony's AntConc 4.3.1 for your operating system.
5. Now, you are ready to start searching for patterns!

❖ Basic Practice:

1. Open AntConc 4.3.1 and upload one to ten corpus files. It is up to you!
2. Select the "Word" tab and then click on "Start" at the bottom.
3. Try to answer the following questions:

WORD COUNT	
A. How many word types are in your corpus?	
B. How many word tokens are in your corpus?	

4. Next, select the "KWIC" tab, type one word from our class list, and click on "Start."
5. Try to answer the following question:

WORD SEARCH	
A. How many 'concordance hits' does this word have?	

6. Using the same word (a.k.a. search term), practice sorting the data 'to the left' and 'to the right.'

7. Try to answer the following questions:

WORD SORT	
A. What is the most frequent word occurring in the first right (1R) position of the search term?	
B. What is its part of speech?	
C. What is the most frequent word occurring in the first left (1L) position of the search term?	
D. What is its part of speech?	
E. How does this fit grammatically with the search term's part of speech?	

8. Next, find an example sentence containing your search term and write it below. You may need to 'expand the context' to 100 or more characters using the "File" tab.

9. Next, find a phrase containing your search term by selecting the "Cluster" or "N-Gram" tabs and then click on "Start." Write the phrase below and try to answer the following question:

A. How many times does it occur?	

10. Look for three more phrases containing your search term. Write the phrases below and try to answer the following questions:

A. How many times does it occur?	
B. How many times does it occur?	
C. How many times does it occur?	

11. Altogether, what grammatical patterns do you see in the phrases? Describe them below:

12. Next, look for other forms of your search term by using the 'wild card' symbol (a.k.a. an asterisk, *). Type part of the word and * with no space. For example, create → create*.

13. Try to answer the following questions:

WILD CARD	
A. What is your search term and the different members of its family?	
B. What is the most frequent word in its family?	
C. What part of speech is it?	

14. Next, return to the "KWIC" tab and select the "Advanced Search" box option. This allows you to search for more terms at the same time.
15. Once the window opens, type in the words *have, get, make, help,* and *let*. Select the "Use search term(s) from list below" box option, then click on "Apply" and "Start."
16. Describe what happens below:

17. Next, find words that 'hang out' with your search term by selecting the "Collocate" tab. Type your search term and click on "Start."
18. Try to answer the following questions:

WORD COLLOCATE	
A. What are the most frequently occurring words at different positions (left and right) of the search terms? Write the words below.	
LEFT:	RIGHT:

19. Lastly, try to use your search term and common collocates in a few sentences of your own.

1.	
2.	
3.	

20. Nicely done!!!😄

Note. Adapted from Green (2019a).

Figure 2. AntConc basics

If working with advanced-level learners, consider using Anthony's (2024b) TagAnt 2.1.1 as well. TagAnt is a freeware, multi-language segmenter, and Part-Of-Speech (POS) tagger. By downloading TagAnt 2.1.1, users can add POS tags to any database, which allows for more in-depth analysis (e.g., adverb + adjective = RB_JJ).

Corpus use in the classroom

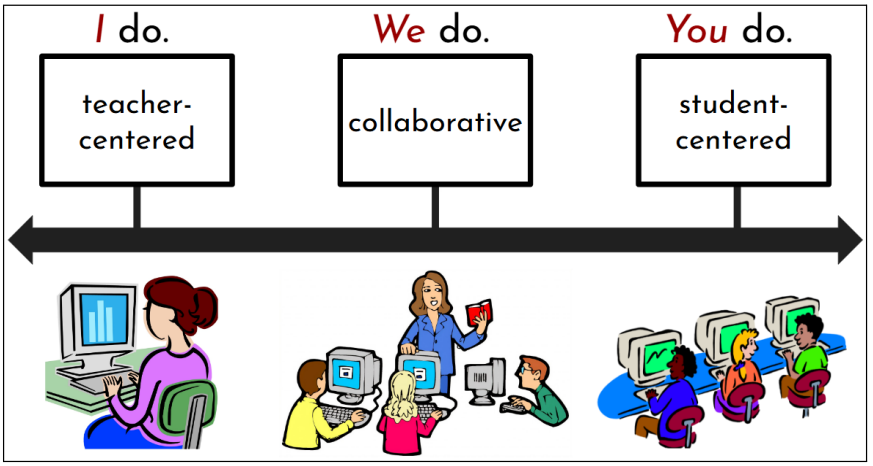
Three simple ways to incorporate corpus technology and familiarize learners with corpus query functions are explained in "*I* do, *We* do, then *You* do" (Nitta, 2022). From experience, this continuum serves as a practical guide for scaffolding forms

of indirect and direct corpus use, resembling teacher-centered, collaborative, and learner-centered tasks (see Figure 3).

First, *I* do: The teacher selects a word or phrase to investigate, retrieves at least ten concordance lines, and guides learners through the generalization process using questions with varying degrees of control (Nitta, 2022; Green, 2019b). For instance, to refine learners' research proposals, my class analyzed the noun/verb *request* using a teacher-generated printout (see Figure 4). This is a form of indirect corpus use, as students do not interact with the corpus directly.

Second, *We* do: the teacher and learners agree on the word or phrase, browse the corpus together, and the teacher helps refine the learners' generalizations. For example, to inform learners' two-sided arguments, my class investigated the collocation and phraseological behavior of near-synonyms *support* and *justify* using COCA. This is a form of direct corpus use, as students are now exposed to raw corpus data (e.g., KWIC, Collocate).

Third, *You* do: learners independently form language questions and browse the corpus while the teacher hardly interferes (Nitta, 2022; Green, 2019b). In a lab setting, one learner may choose to investigate near-synonymy using WebParaNews, while another examines relative clauses using COCA. This is the final form of direct corpus use, as learners are browsing corpora independently to answer their language questions.



Note. Adapted from Nitta (2022), Hinkel & Green (2022), and Sripicharn (2003).

Figure 3. Corpus use continuum

NAME: _____

CLASS: _____

The Noun/Verb "Request"

❖ **Instructions:**

"Request" has both noun and verb forms and frequently occurs with other words in various texts. In this task, examine the following example sentences with "request" from COCA. What patterns do you notice? Some guiding questions have been provided to help you guess the patterns. You may work with a partner or in a small group.

❖ **Example Sentences:**

1. The trial court may request permission to address the petition… <WEB, 2012>
2. Any State Party may request assistance relating to medical treatment… <WEB, 2012>
3. The Daily News Leader requested information from the Shenandoah Valley Juvenile Center. <NEWS, 2019>
4. The investigations were requested by Washington State President Elson Floyd. <WEB, 2012>
5. The hearing was requested by the Illinois Press Association… <WEB, 2012>
6. The judge signed off on the doctor's request. <ACAD, 2019>
7. The woman's request for a temporary restraining order was approved. <NEWS, 2019>
8. O'Neill turned down the president's request. <NEWS, 2019>
9. Material will be removed upon request of the copyright owner. <WEB, 2012>
10. For a complete list of co-authors, the full paper and editorial is available upon request. <BLOG, 2012>

❖ **Guiding Questions:**

A. Can you find examples of "request" as a verb? Write the number of each sentence.	
B. Can you find examples of "request" as a noun? Write the number of each sentence.	
C. What are the most frequent words that occur before or after "request" as a verb?	
D. What are the most frequent words that occur before or after "request" as a noun?	
E. What else do you notice about the use of "request" in various texts?	

Note. Adapted from Nitta (2022).

Figure 4. Teacher-centered task

Application of Illustration-Induction-Interaction

Following the Illustration-Induction-Interaction (I-I-I) model (McCarthy & Carter, 1995), teachers can further develop learners' corpus query techniques by sequencing tasks that lead to a culmination activity or formative assessment. For instance, in an advanced writing course, I implemented the I-I-I model for a vocabulary lesson (Nitta, 2023). The objective of the corpus-informed lesson was to differentiate between noun near-synonyms *attribute, characteristic*, and *trait*. Using a teacher-generated printout like Figure 4, learners were given concordance lines (for Illustration)

with guiding questions (for Induction). Then, for Induction and Interaction, learners made semantic associations by categorizing the collocates into different meaning groups (see Figure 5). Learners were encouraged to use corpus tools (e.g., Collocate) to verify, refute, and refine their hypotheses about the target words (Timmis, 2015; Nelson, 2006). Finally, for further Interaction, the formative assessment instructed learners to complete a writing task in which they must accurately use the target words. The I-I-I model led to opportunistic decisions, memorable teaching moments, and empowered learners to differentiate between near-synonyms independently.

NAME: _____

CLASS: _____

"Attribute," "Characteristic," & "Trait"

❖ **Instructions:**

"Attribute," "characteristic," and "trait" have different semantic associations in academic texts. In this task, categorize the following collocates of "attribute," "characteristic," and "trait" into the meaning groups provided. What patterns do you notice? Use corpus tools to explore these associations. You may work with a partner or in a small group.

❖ **Semantic Associations:**

COLLOCATES
attribute of security
attribute of a good or service
characteristics of students
characteristics of participants
traits of substance abuse
traits of autism

MEANING GROUPS		
standard or quality	demographic, species, or property	cognition, behavior, or genetics

Note. Adapted from Nitta (2023).

Figure 5. Induction-interaction task

Conclusion

Corpus technology has gained notable attention over the past decade for its databases, tools, and uses. In the age of AI, many types of corpora are increasingly accessible, with software like AntConc providing a friendly introduction to corpus basics. By implementing a variety of structured – isolated or sequenced – tasks and applications like I-I-I, teachers can effectively use corpus technology to guide learners toward independent learning.

JALTCALL Trends
Vol. 1
2024 JALTCALL Conference Papers Special Issue

References

Anthony, L. (2024a). AntConc (Version 4.3.1) [Computer Software]. Tokyo, Japan: Waseda University. https://www.laurenceanthony.net/software

Anthony, L. (2024b). TagAnt (Version 2.1.1) [Computer Software]. Tokyo, Japan: Waseda University. Available online at https://www.laurenceanthony.net/software

Anthony, L. & Chujo, K. (n.d.). WebParaNews [Website]. Available online at https://www.antlabsolutions.com/webparanews/

Biber, D. (1996). Investigating language use through corpus-based analyses of association patterns. *International Journal of Corpus Linguistics, 1*(2), 171–197.

Biber, D., Reppen. R. & Conrad, S. (2007). *Corpus linguistics: Investigating language structure and use.* Cambridge University Press.

Boulton, A. (2010). Data-driven learning: Taking the computer out of the equation. *Language learning, 60*(3), 534–572.

corpus. 2024. In *Merriam-Webster.com.* Retrieved May 1, 2024, from https://www.merriam-webster.com/dictionary/corpus

Davies, M. (2008-). The Corpus of Contemporary American English (COCA) [Website]. Available online at https://corpus.byu.edu/coca/

English Language Institute. (1997–2002). The Michigan Corpus of Academic Spoken English (MICASE) [Website]. Available online at https://quod.lib.umich.edu/m/micase/

English Language Institute. (2009). The Michigan Corpus of Upper-Level Student Papers (MICUSP) [Website]. Available online at https://elicorpora.info/main

Friginal, E. (2018). Corpus linguistics for English teachers: New tools, online resources, and classroom activities. Routledge.

Green, B. A. (2019a). *Class handout on AntConc basic use.* Unpublished class handout, LING 321, Brigham Young University-Hawai'i.

Green, B. A. (2019b). Teaching and assessing multiword expressions using an open educational resources academic English corpus. In E. Hinkel (Ed.), *Teaching essential units of language: Beyond single-word vocabulary* (pp. 134–164). Routledge.

Hinkel, E. & Green, B. A. (2022, March 24). Pedagogical approaches to teaching and learning multiword expressions. TESOL International Convention. Pittsburgh, Pennsylvania.

Johns, T. (1991). Should you be persuaded: Two samples of data-driven learning materials. *English Language Research Journal, 4,* 1–16.

McCarthy, M. & Carter, R. (1995). Spoken grammar: What is it and how can we teach it? *ELT Journal, 49*(3), 207–218.

Nelson, M. (2006). Semantic associations in Business English: A corpus-based analysis. *English for Specific Purposes, 25*(2), 217–234.
Nitta, S. M. M. (2022). Voices from the field: Three ways to incorporate corpus technology in the ESL classroom. *TESOL Working Paper Series, 20,* 53–57.
Nitta, S. M. M. (2023). A corpus analysis of plesionyms attribute, characteristic, and trait in Academic English. *TESOL Working Paper Series, 21,* 80–120.
Sripicharn, P. (2003, December 16). Implementing collaborative concordancing between teacher and learners in the writing class. Paper presented at the 5th CULI International Conference, Bangkok, Thailand.
Timmis, I. (2015). *Corpus linguistics for ELT: Research and practice.* Routledge.

JALTCALL Trends
Vol. 1
2024 JALTCALL Conference Papers
Special Issue

Author biodata

Sadie Nitta holds an MA in Teaching English to Speakers of Other Languages from Hawai'i Pacific University. She has taught English at universities in Hawai'i and Japan as well as contributed to various corpus projects. Her main research interests are corpus linguistics, sociolinguistics, and peace linguistics.