

Using Learner Corpus Data for Language Teaching: Introduction to Antconc

SLAT roundtable - Spring 2018

AntConc is a free downloadable program, constantly being updated:

<http://www.laurenceanthony.net/software/antconc/>

- It was created by Laurence Anthony → if you have questions, email him!
- But first, watch these tutorials: [Getting started](#); [Basic features](#); [Advanced features](#);
- [Concordance plot tool](#); [File view tool](#); [Clusters tool](#); [N-grams tool](#); [Collocates tool](#);
- [Word list tool](#); [Keyword list tool](#); [Working with tagged data](#)

AntConc can be used with any group of texts. However, your files need to be in .txt format or .xml format. You can use the free programs [Zilla Word-to-Text Converter](#) or [AntFileConverter](#) to convert word files and other common file formats to plain text.

Useful terminology:

- **Keyword:** words that occur more often in one text or corpus than in another. This is useful because it helps us to identify words that are more associated with one particular group of texts in relation to another.
- **Cluster/n-gram:** also called lexical bundles, groups of words that appear frequently in sequence in a text.
- **Collocation:** collocations are words that frequently appear together; when two/more words collocate with each other, it means they appear together in language use, e.g., *tea* collocates with *black*, or *green*; *pizza* collocates with *pepperoni*.
- **Concordance:** words to the right and to the left from a word you search for; programs usually allow you to include a specific number of words to the left and/or to the right of the search word.
- **KWIC (Key Word in Context):** is what you see in the concordance line when the searched word/phrase (or key word) is put in the middle to see the left and the right concordances.
- **Range:** the number of texts that the word is found in.

Arizona Second Language Writing Corpus (ASLW)

ASLW is a corpus of first year writing texts from UA's English 106, 107 and 108 courses. The corpus consists of 14 written assignments each semester and multiple drafts of each assignment. The processed corpus contains over 700,000 words (Fall 2016 - Spring 2017) and is available as plain text for University of Arizona students and researchers. The corpus continues to grow each semester.

Spanish Learner Language Oral Corpora (SPLLOC)

SPLLOC is a corpus of spoken Spanish produced by 60 L1 English speakers, from beginners to advanced level, who have learned L2 Spanish in educational contexts within the UK. For comparison purposes, native speakers were also recorded undertaking the same tasks. The resulting database of L2 Spanish contains digital soundfiles of learner speech, in varying genres, accompanied by transcripts in CHILDES format. We have reformatted the transcripts to be compatible with AntConc.

How to cite this: Crow Team (2018). Using Learner Corpus Data for Language Teaching: Introduction to Antconc [Handout]. SLAT Roundtable, University of Arizona, Tucson, AZ.

Practice using AntConc - Activities

Before doing any of the activities below, make sure you have your tags hidden:

1. Go to “Global Setting,” choose the “Tags” category.
2. Click on “Hide Tags.”
3. Click “Apply.”

**** Activities with Learner Data in English ****

Activity 1: Word list

1. Click on “File” and then choose “Open Dir...”, choose “Desktop” then choose “Roundtable data” (click arrow on a Mac). Then, click on “English,” and then click on “Narrative,” and Press “OK.”
2. Go to “Word List” tab (at the top of the window), and click “sort by Freq.”
3. Click “Start.”
4. What are the top 5 function words? What about the top 5 content words?
5. Let’s compare these results with the word list obtained from texts in a different genre. To do this, first click on “File” and then choose “Clear All Tools and Files”.
6. Click on “File” and then choose “Open Dir...”, choose “Desktop” then click “roundtable data,” (click arrow on a Mac), click on “English,” then choose “Rhetorical Analysis,” and Press “Open.”
7. Follow steps 2 to 4 above.
8. What are the main differences in results? Why?

Activity 2: Clusters function (What are the most common lexical bundles/n-grams in the argumentative task?)

1. Keep the same files you have already loaded in step 6 in Activity 1 (“Rhetorical Analysis”) - In other words. Don’t do anything. Go to the next step.
2. Go to the “Clusters” tab.
3. Check the “N-Grams” box under Search Term. Change the minimum size to “3” and maximum size to “4.” Click “Start.”
4. Under “Sort by” choose “Sort by Range.”
5. Click “Start.”
6. Note the most common n-grams.

Activity 3: Concordance function (what rhetorical function do the n-grams have)?

1. Using the same files from Activity 2 (step 1 above)
2. Go to the “Concordance” tab.
3. Type “because” in the Search Term box.
4. Click “Start.”
5. Then under “Kwic Sort” keep Level 1 on 1R, and deselect Levels 2 and 3. Click “Sort.”
6. Scroll down to see the environment in which “because” is used.
7. Type “if” in the Search Term box, follow steps 4-6.
8. Once finished, click “file” and then “Clear all tools and files” to clear your AntConc program for the next activity. (This sometimes causes AntConc to act weirdly. If this happens, restart AntConc).

How to cite this: Crow Team (2018). Using Learner Corpus Data for Language Teaching: Introduction to Antconc [Handout]. SLAT Roundtable, University of Arizona, Tucson, AZ.

****** Activities with Learner Data in Spanish ******

Activity 1: Word list

1. Click on “File” and then choose “Open Dir...”, choose the “Desktop” folder then click on “Roundtable data” (click arrow on a Mac). Click on “Spanish,” and then click on “Modern Times Narrative,” and then just select (don’t click on it twice) “undergraduate”. Press “Open.”
2. Go to “Word List” tab (at the top of the window), and click “sort by Freq.”
3. Click “Start.”
4. What are the top 5 function words? What about the top 5 content words?
5. Let’s compare these results with the word list obtained from texts in a different genre. To do this, first click on “File” and then choose “Clear All Tools and Files”.
6. Click on “File” and then choose “Open Dir...”, choose the “Desktop” folder then click on “Roundtable data” (click arrow on a Mac). Click on “Spanish,” and then click on “Photo Interview,” and then just select (don’t click on it twice) “undergraduate”. Press “Open.”
7. Follow steps 2 to 4 above.
8. What are the main differences in results? Why?

Activity 2: Clusters function (What are the most common lexical bundles/n-grams in the argumentative task?)

1. Keep the same files you have already loaded in step 6 in Activity 1 (“Photo Interview”) - In other words. Don’t do anything. Go to the next step.
2. Go to the “Clusters” tab.
3. Check the “N-Grams” box under Search Term. Change the minimum size to “2” and maximum size to “3.” Click “Start.”
4. Under “Sort by” choose “Sort by Range.”
5. Click “Start.”
6. What are the most common clusters?

Activity 3: Concordance function (what rhetorical function do the n-grams have)?

1. Keep the same files from Activity 2.
2. Go to the “Concordance” tab.
3. Type “después” in the Search Term box (without the quotation marks).
4. Click “Start.”
5. Then under “Kwic Sort” keep Level 1 on 1R, and deselect Levels 2 and 3. Click “Sort.”
6. Scroll down to see the environment in which “después” is used.
7. Type “esta*” or “habí*” in the Search Term box, follow steps 4-6.
8. What are the most common verb forms used by learners?

* * * *

What if I want to export these examples for my students (or for further analysis)?

1. Go to File, then choose “Save Output to Text File.” Save file.
2. Open Excel and open the txt document. Click through the options to format the txt file as an Excel file.

How to cite this: Crow Team (2018). Using Learner Corpus Data for Language Teaching: Introduction to Antconc [Handout]. SLAT Roundtable, University of Arizona, Tucson, AZ.

3. Note that you cannot directly export from the Concordance Plot function. For this, you can use a screenshot instead.