Introduction

Corpus analysis tools can be used to create personalized vocabulary lists for students to use during pre-reading activities. Nearly any general-purpose corpus analysis tool that can create frequency lists can be used to do this.

The basic steps are to first run an electronic version of the reading text through a word frequency list generator. Then, present the resulting list to students. Students browse the list and mark any words they do not know. Every student now has a personalized vocabulary list that they can study before reading the full text.

Further optional steps can be added to the base of this process that will result in more finely-tailored lists.

This poster illustrates the steps for creating vocabulary lists for a text that allow teachers and students to approximate what percent of the text’s words students can understand before reading, or what words should be studied before reading.

Tools and Terminology

Antconc: A general-purpose corpus analysis tool. The process described in this poster is not limited to Antconc.

Frequency list: A list of how many times each word in a text appears in that text.

Lemma list: A list that can be added to a frequency list generator that will count various forms of a headword together (i.e. counting ‘am’, ‘is’, and ‘are’ together under the headword ‘be’).

Stop list: A list that will omit from a frequency list words specified by the user (i.e. function words, very high-frequency words, or previously studied words).

Word token: An individual word identified as such by the corpus analysis software.

Steps and Options

1. Open a .txt version of the class text in Antconc.
2. Choose the Word List tab and press [Start].
3. Record how many word tokens are in the text.
4. Optional: Open the [Tool Preferences] menu, and select [Word List].
5. Optional: Load a lemma list and/or a stop list, and press [Apply].
7. Copy the word list and note the frequency of each word/lemma.
8. Make a list of the words and their frequencies for students.
9. Students mark words they do not know.
10. Add the frequency counts of unknown words and divide by the TOTAL number of word tokens in the text.
11. Write as a percent, and subtract from 100.
12. The result approximates the percent of words in the text a student will know.

Limitations

The basic version of the process described here does not take into account forms and features such as phrasal verbs, polysemy, homography, etc., which may lead to misunderstandings in the case of, for example, ‘wind’.

Also, this process looks at words outside of their context(s). It may be that students will mark a word as ‘unknown’ which they would in fact have no problem with when seeing it in context.

Background and References

