Research Tools: Wordlists (Offsite)

What can you do with the Wordlists page?

- Get frequencies or concordance lines including all the words from a number of online wordlists;
- Get summary information for all the wordlists;
- View the wordlist for the currently selected corpus.

When you are connected to the server, you can also use the same wordlists to get frequency information for your own DIY Corpora. See **tPM Help 009 DIY Text Tools** for further details.

How does it work?

On the server, a number of wordlists are available. By selecting one of these, you can receive frequency information or concordance lines for all the words on the list.

Further details

The Wordlists page on the Research Tools tab allows you to retrieve frequencies and concordance lines using a set of word lists, or to generate a word list for the currently selected online corpus.

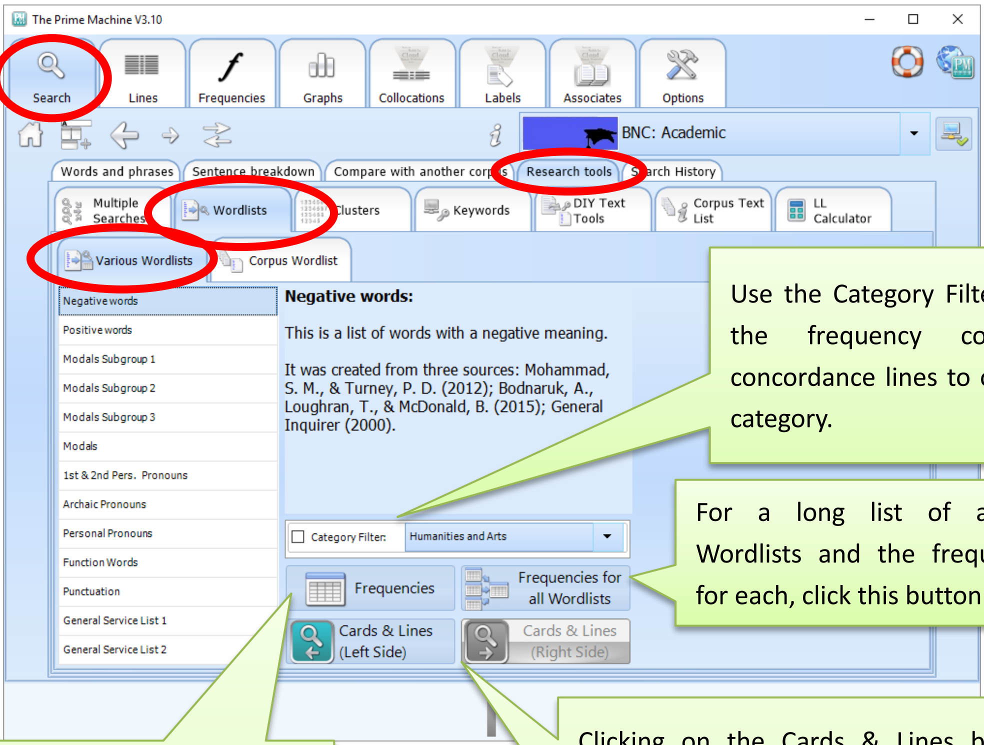
Various Wordlists

There are a number of wordlists available, ranging from simple lists of specific modal verbs, to wordlists often used to evaluate vocabulary difficulty in language teaching materials.

Staff and students at XJTLU can connect to the server using the tPM Home Network, and gain access to a wider range of wordlists. They should refer to **tPM Help 006b Wordlists on tPM Home Network**, which can be found on the XJTLU Staff and Students pages of www.theprimemachine.com.

You can get information about each wordlist by selecting it from the list. Further details are for wordlists available offsite are provided here.

The Prime Machine does not give you access to these specific lists; it simply provides data from the online corpora matching words on these lists. For the wordlists, see the relevant references.



Negative words:

This is a list of words with a negative meaning.

It was created from three sources: Mohammad, S. M., & Turney, P. D. (2012); Bodnaruk, A., Loughran, T., & McDonald, B. (2015); General Inquirer (2000).

Use the Category Filter to limit the frequency counts or concordance lines to one major category.

For a long list of all the Wordlists and the frequencies for each, click this button.

The Frequencies button retrieves a list of words from the wordlist in the currently selected corpus, along with their frequencies. Up to 1,000 results will be retrieved, sorted in descending order of frequency.

Clicking on the Cards & Lines buttons will retrieve concordance lines containing any of the words from the lists.

For very long wordlists, this feature is not available.

Wordlist	Items	Details
Negative words	2,495	This is a list of words with a negative meaning. It was created from three sources: • Mohammad, S. M., & Turney, P. D. (2012); • Bodnaruk, A., Loughran, T., & McDonald, B. (2015); • General Inquirer (2000).
Positive words	618	This is a list of words with a positive meaning. It was created from three sources: • Mohammad, S. M., & Turney, P. D. (2012); • Bodnaruk, A., Loughran, T., & McDonald, B. (2015); • General Inquirer (2000).
Modals Subgroup 1	6	This is the first group of modals: *will, would* and *shall*. It also includes contracted forms.
Modals Subgroup 2	4	This is the second group of modals: *can, could, may* and *might*.
Modals Subgroup 3	4	This is the third group of modals: *must, should, need* and *ought*.
Modals	15	These are all three subgroups: *will, would* and *shall*; *can, could, may* and *might*; *must, should, need* and *ought*. Some contracted forms are also included. The groupings are based on Biber et al. (1999).
1st & 2nd Pers. Pronouns	14	These are the personal pronouns for first and second persons (*I, you, we*, etc.).
Archaic Pronouns	6	These are some archaic pronouns (*thy, thee*, etc.) used in older English texts.
Personal Pronouns	28	These are the list of modern personal pronouns (first, second and third persons).
Function Words	327	This is a subset of the General Service List 1 (see below) which contains grammatical words.
Punctuation	21	This is a list of frequently used punctuation marks for English.

Wordlist	Items	Details
General Service List 1	4,114	This is a list of the most frequent 1,000 word families for English, based on the General Service List (West 1953). The word families list was downloaded from https://www.lextutor.ca/freq/lists_download/ (Cobb 2000).
General Service List 2	3,708	This is a list of the second thousand most frequent word families for English, based on the General Service List (West 1953). The word families list was downloaded from https://www.lextutor.ca/freq/lists_download/ (Cobb 2000).
Academic Word List	3,082	This is the Academic Word List (Coxhead 2000), containing the most frequent word families outside the first 2,000 words from GSL1+GSL2, occurring in academic texts across disciplines. The word families list was downloaded from https://www.lextutor.ca/freq/lists_download/ (Cobb 2000).

Other resources are only available when connected to tPM Home Network.

> ⚠ **Wordlist data is based on simple text matches**
> No attempt has been made to distinguish between different words or senses of words with the same spellings.
> For example: *may* will match the modal verb use, the name and the month of the year.

Additional notes on the Positive / Negative wordlists

All the positive and negative words from Bodnaruk, Loughran and McDonald were included. For words from Mohammad and Turney and the General Inquirer to be included:

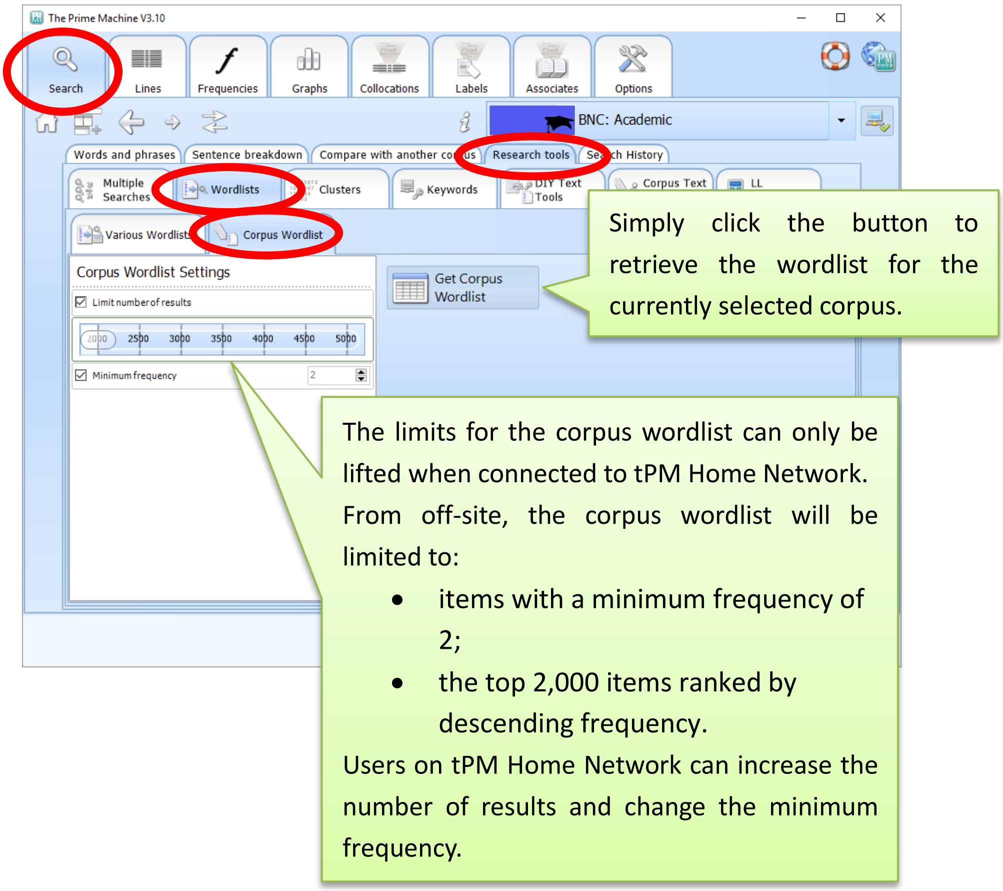
1) They must have occurred on lists from both sources;
2) They must not also occur in the opposite list;
3) On the General Inquirer list, they must not have the # symbol (meaning it has several meanings)

For more information about the Loughran and McDonald Word Lists, see https://www3.nd.edu/~mcdonald/Word_Lists.html

Corpus Wordlist

This page allows you to get a list of words and frequencies for the currently selected online corpus. If you are connected to tPM Home Network, you will be able to adjust the number of results and the minimum frequencies.

The table of results shows the raw frequency, the normalized frequency per million words, the proportion of occurrences across each of the major categories and a note on the number of texts containing the item.



☑ Limit number of results

☑ Minimum frequency 2

Simply click the button to retrieve the wordlist for the currently selected corpus.

The limits for the corpus wordlist can only be lifted when connected to tPM Home Network. From off-site, the corpus wordlist will be limited to:

- items with a minimum frequency of 2;
- the top 2,000 items ranked by descending frequency.

Users on tPM Home Network can increase the number of results and change the minimum frequency.

References

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Harlow: Longman.

Bodnaruk, A., Loughran, T., & McDonald, B. (2015). Using 10-K Text to Gauge Financial Constraints. *Journal of Financial & Quantitative Analysis, 50*(4), 623-646. doi: 10.1017/s0022109015000411

Cobb, T. (2000). The Compleat Lexical Tutor, from http://www.lextutor.ca

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213-238.

GI. (2000). General Inquirer: URL: http://www.wjh.harvard.edu/~inquirer/homecat.htm.

Mohammad, S. M., & Turney, P. D. (2012). Crowdsourcing a word–emotion association lexicon. *Computational Intelligence, 59*.

Nation, I. S. P., *Paul Nation's web pages, containing resources and links*. Available from https://www.victoria.ac.nz/lals/about/staff/paul-nation

West, M. (1953). *A General Service List of English Words*. London: Longman, Green and Co.

Support

The Prime Machine is still undergoing development.
For further information see http://help.theprimemachine.com

Last Updated: Friday, April 13, 2018