

8 Ngram tool

The Ngram tool allows in-depth analysis of frequencies of n-grams (bigrams, trigrams etc.), which could be defined as contiguous combinations types, lemmas and POS. The tool also produces key ngrams by comparing two corpora using a technique similar to keywords.

It can be used, for example, to:

- Identify n-grams, lexical bundles and p-frames (also skip grams)
- Compute frequency and dispersion measures for ngram types, lemmas and POS tags.
- Visualize frequency and dispersion of ngrams in corpora.
- Compare ngrams in two corpora using the keyword technique.
- Visualize key ngrams.

8.1 Visual summary

The screenshot displays the #LancsBox v 4.0 Ngram tool interface. The top menu bar includes options like KWIC, GraphColl, Whelk, Words, Ngrams, and Text. The main window is divided into two panes. The left pane shows a table of n-grams for the 'L-O-B' corpus, with columns for Type, Frequency, Dispersion, and Grams. A callout box points to the table header with the text 'Right-click on the table header to activate filter.' The right pane shows two corpus icons, 'L-O-B' (pink) and 'Brown' (black), with callout boxes: 'Drag corpora together to produce key ngrams.' and 'Left-double-click on the corpus to see its internal structure.' Below the corpus icons, a callout box points to the 'L-O-B' icon with the text 'Right-click on the corpus to see corpus statistics.' The bottom of the interface shows a status bar with 'Filtering complete' and navigation icons.

Left: Creating frequency lists, computing dispersion and key ngrams.

Right: Visualizing frequencies, dispersions and key ngrams.

► Did you know?

Multi-word expressions are extremely important when describing language. There are different terms to describe multi-word expressions such as collocations (Brezina et al. 2015; Gablasova et al. 2017), n-grams, lexical bundles and p-frames. While collocations, which are identified in the GraphColl module, typically represent non-contiguous expressions, the n-gram type multi-word expressions represent contiguous lexico-grammatical patterns. They are defined as follows.

- n-gram: a sequence of n types, lemmas, POS from a text or corpus.
- lexical bundle: an ngram with certain frequency and distributional (dispersion) properties, e.g. relative freq. 10 per million and range > 5.
- p-frame (also skip gram): an n-gram that allows for variability at one or more positions such as *it would be * to*.

All these types of multi-word expressions can be identified using the Ngram tool in #LancsBox.

Brezina, V., McEnery, T. & Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics*, 20(2), 139-173.

Gablasova, D., Brezina, V., & McEnery, T. (2017). Collocations in corpus-based language learning research: Identifying, comparing, and interpreting the evidence. *Language Learning*, 67 (S1), 155–179.