DH2019
UTRECHT UNIVERSITY 8-12 JULY

Utrecht University

Workshop session 3: Introducing linguistic analysis of text: free tools

Suggestions including Feedback from SketchEngine team

Jeannine Beeken, 11 July 2019

The shared folder contains

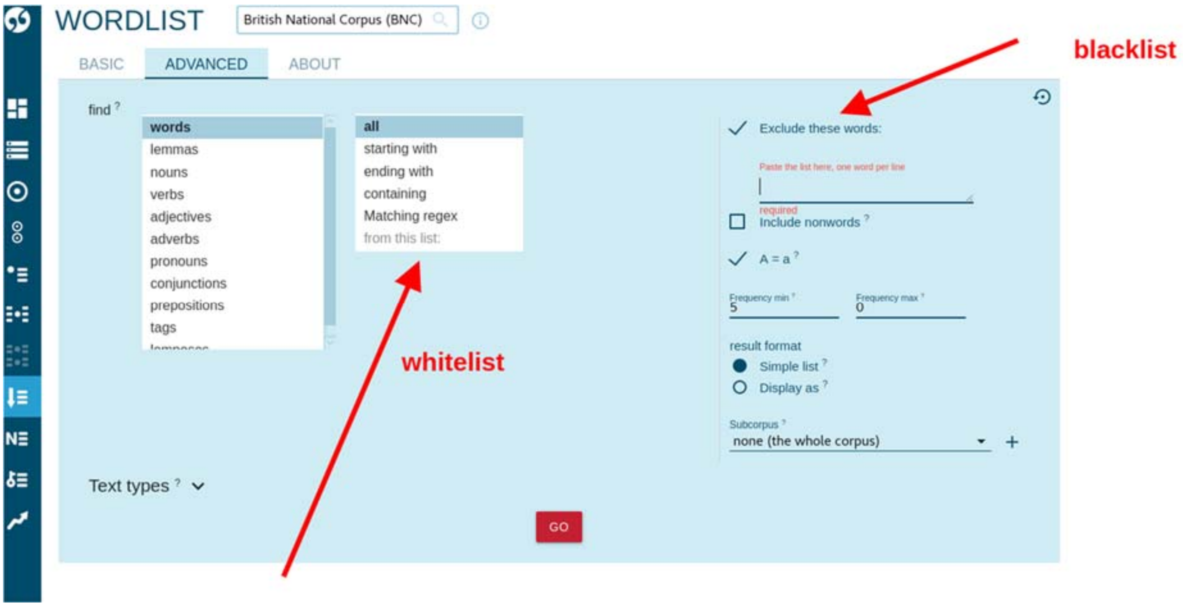
1. A document with linguistic *background information*, containing *terminology* and many *clarifying examples*.

2. A document with the *tasks* re *SketchEngine Basic*, providing guidelines concerning 1) how to create a corpus and subcorpora, 2) how to explore corpora based on semantic differences and similarities.
   Section 4 of the document contains subsections a-h. These subsections explore and clarify the different *functionalities* of SketchEngine on a Basic level. They also contain questions (in italic) illustrating the *usefulness and usage of the tool*. For example:

   a. use Keywords of a corpus to enable you to locate the interviewees according to geographic location, language register (dialect, slang, archaic, obsolete etc.).
   b. use Keywords of the subcorpora to find differences in time and geographical locations, religion, origin, race, gender etc.
   c. N-Grams: skipped
   d. Wordlist: the importance of words vs. lemmas in order to cluster the frequencies enabling you to generate lists according to meaning (to be; child) instead of lists according to form (am, is, was, were etc.; child, children).
   e. Concordance: WSD (word sense disambiguation) and POS (part-of-speech tagging) in order to assign the correct meaning as intended by the interviewee and in order to interpret collocations and idioms in a correct semantic way, i.e. the meaning of the total (for example idioms) is not the same as the sum of the meanings of the separate words. Examples: watch (‘clock’ or ‘see’), ring (‘wedding ring’ or ‘phone someone’ or ‘the sound of bells’), ‘all went black for me’, ‘black sheep of the family’, ‘black death’ etc.
   f. Thesaurus enables you to generate a corpus-dependent semantic word cloud of similar according to word class (POS: for example the adjective ‘black’ vs. the noun ‘black’)
   g. Word Sketch enables you to see the concordances, co-occurrences and context of a word or multi-word. Examples are the WordSketches of ‘house’, ‘dead’ etc.
   h. WordSketch difference enables you to compare the ‘surroundings’ of synonyms or similar words. This assists in analysing content using similar concepts, for example ‘big’ and ‘great, ‘see’ and ‘watch’ or ‘small’ and little’.

3. Your questions and feedback have been forwarded to the SketchEngine team. Their answers have been added to the questions below:
    a. Q: Where can I find the standard Word lists annex frequencies?
    A: We have still the standard Word list function, perhaps some users might be confused by missing frequencies of words which can be displayed via View option.

    b. Q: Where can I find the blacklisting (stopwords) feature in Basic?
    A: Via Advanced tab in Word list feature, please see the attached screenshot. It is a slightly different logic from previous interface, but it is working same as before.



WORDLIST British National Corpus (BNC)

find: words / all

☑ Exclude these words:

Paste the list here, one word per line (required)

□ Include nonwords

☑ A = a

Frequency min: 5

Frequency max: 0

result format: ● Simple list ○ Display as

Subcorpus: none (the whole corpus)

    c. Q: How are the Keywords sorted?
    A: The keywords are sorted by the typicality score which is counted via simple math method. You find how this score is counted here

    d. Q: Is it possible to select subcorpora in the box next to the DASHBOARD?
    A: You can select subcorpora in each tool separately on Advanced tab, e.g. Concordance, word sketch, word list, etc. Only you need to open the tool and switch to Advanced tab (the Basic tab shows only basic options, not selecting subcorpora).

    e. Q: What is the logic behind the sorting of the Sketch results?
    A: Word Sketch and word sketch difference shows results sorted by the logDice score which is statistic and mathematical measure (not only based on raw frequency). Probably, the user displayed only frequency via option, but you can also turn on this logDice score. Of course, you can sort the results by raw frequency.