DH2019
UTRECHT UNIVERSITY 8-12 JULY

Utrecht University

Workshop: The Case of Interview Data, Tuesday 9 July, 09:00-13:00

Session: Introducing linguistic analysis of text: free tools

Workshop Tasks for SketchEngine: a linguistic tool for exploring corpora, word sketches, thesauri, concordances, wordlists, n-grams and keywords

User manual: https://www.sketchengine.eu/user-guide/user-manual/

Jeannine Beeken, UK Data Archive, University of ESSEX

1. Go to https://auth.sketchengine.eu/#login. Choose either Sign up for a free 30-day trial or choose 'Institutional login'. Check whether your institution is part of the list (ELEXIS project) If so, you will get direct access using the Institutional login where you can type your login name and password.

2. The DASHBOARD is displayed. Familiarize yourself with the layout and meaning of the symbols.

3. First you need to create a CORPUS OF TEXTS.
   By default, for English, the English Web 2015 corpus should have been chosen. If not, please click on the search button (in the box next to DASHBOARD) and click on English Web 2015. Press on 'CORPUS INFO' to get general info, counts, common tags, lempos suffixes (word class of the lemma) and lexicon sizes together with some subcorpora statistics. Close the window.

   a. Create your own corpus and subcorpora (use the four UKDA txt documents).
      Click on NEW CORPUS. Follow the instruction bar.

   b. Name your corpus, choose the language (i.e. English) and fill out the description.

   c. Check the available features by clicking on the arrow. Press next.

   d. On the DASHBOARD, look at RECENTLY USED CORPORA and choose your corpus.

   e. You will be notified that your corpus is empty and that you need the ADD TEXTS (either web texts or your own). Click on ADD TEXTS and choose 'I have my own texts'.

   f. Click on 'Choose a file' and upload the four UKDA texts (interviews) all at once. The CORPUS CONTENT displays the number of words.

   g. Click on Next and then Compile.
      Click on GET TO KNOW YOUR CORPUS: extract keywords & terms (single-words/multi-

words) or details & statistics. The four documents you uploaded form 1 corpus. But you can also create subcorpora in order to compare of select specific documents.

h. Left-hand side: click on DASHBOARD and click on MANAGE CORPUS.
Click on 'Subcorpora' and then on 'Create Subcorpora'. Give the subcorpus a name, for example 'interview1' and select a File name of the list of three documents. Save by clicking on the Save icon at the bottom, right-hand side.
Repeat for the three other documents.

i. Go to the DASHBOARD and choose your corpus from the list of RECENTLY USED CORPORA or MY CORPORA.

4. Now you can start EXPLORING THE CORPUS AND SUBCORPORA you created.

a. Use the menu on the lhs: start from the bottom and choose Keywords (key symbol).
Click on GO and you will see a list of single-words and multi-words.
Scroll down and extend the list to 200 rows per page (at the bottom)
What are the most frequent single-words of the Focus corpus (your corpus)?
Are there any words that you would mark as 'regional, archaic/obsolete, geographic'?
Scroll down in multi-words: does for example 'big', 'little' or 'bottom' have exactly the same meaning in all of the multi-words?
Is 'awful' used as a negative indicator or a positive one?

b. Optional: go back to Keywords (lhs menu) and select ADVANCED.
Select one of the four subcorpora (documents in this case), for example 'interview 1' and click on Go.
You can change your criteria by clicking on the 'Change criteria' button on the top rhs (next to the download button). Extend the list of results to 200.
Can you spot any different locations, words, unknown words, time/historical indications?
Compare the keywords of interview1 with the keywords of interview2? Are there any important differences re location, time, religion, origin, words used?

c. N-Grams. Let us skip this functionality for the time being.

d. Wordlist. In BASIC, find all the words.
Look for all nouns, verbs or adjectives.
Now change to lemmas.
What are the words with the highest frequency? What is the difference between a word and a lemma and why could this information be useful?

e. Concordance. This is similar to KWiC (KeyWords in Context).
The options BASIC and ADVANCED will be explored.
Choose a text type: either the whole corpus or one of the subcorpora (for the latter choose a File name).
Type in a Simple search term, for example 'pretty', 'watch', 'ring', 'clean' or 'well'.
Do the concordances help in defining and separating their meaning and use? Can you detect any collocations/idioms? (for example 'watch football' not 'see football' etc.)

f. Thesaurus contains synonyms and similar words belonging to the same semantic cluster.
In BASIC search, type a search term such as ‘black’, ‘sister’, ‘house’, ‘home’, ‘man’ or ‘woman’.
Click on the Show visualisation icon (circle of dots) on the right hand side to see a semantic cloud.
Is this the cloud you expected? If not, why not?

g. Word Sketch is a useful novelty offered by the WordSketch Engine. Word sketches can be seen as collocations and word combinations.
Perform a BASIC search for ‘black’ or ‘pretty’ and explore the ‘as adjective, as adverb’ etc. options.
Search for ‘table’. Click on the three dots (…) of ‘wee’ and hover over the options. Choose for example Thesaurus (not the ‘wee’ as used in different parts of the UK others than Scotland 💻).
Search for ‘time’. Click on the ‘… ‘ next to ‘hard’ and click on ‘Concordance’.
Explore for yourself searching for the word sketches of for example ‘house’, ‘dead’, ‘come’ (scroll down).
How can this functionality help you in your research? Is it about the content and or use of language in different circumstances or by different speakers? Or rather, is it more about the peculiarities of the language itself?

h. Word Sketch Difference is designed for making comparisons by contrasting collocations.
Go to BASIC and search for a ‘First lemma’ and a ‘Second lemma’. For example ‘see’ and ‘watch’. Other examples are ‘big’ and ‘great’, or ‘small’ and ‘little’..
What do you learn by using this tool? When do you use which variant/synonym? Could it be speaker, register or situation dependent?