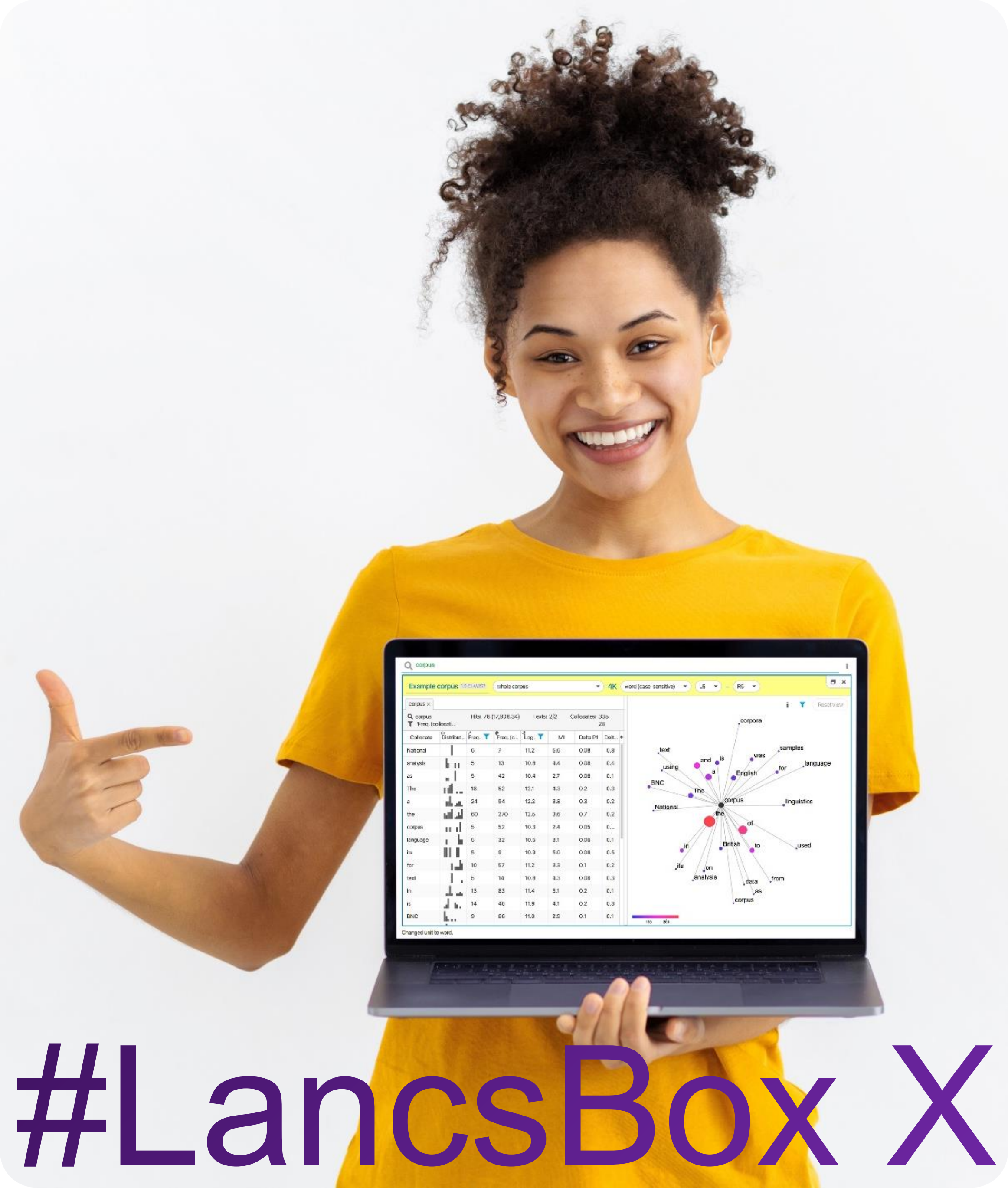


#LancsBox X

Words

Task 1 Understanding frequency

In this task, you will explore the information about the frequency of words. Go to the Words tool in #LancsBox X and select the BNC2014 corpus (whole corpus). Keep the unit as 'word (lowercase)' and find the following information:

1. Number of words in the wordlist __________858,450______

2. The most frequent word in the wordlist is ________the_________

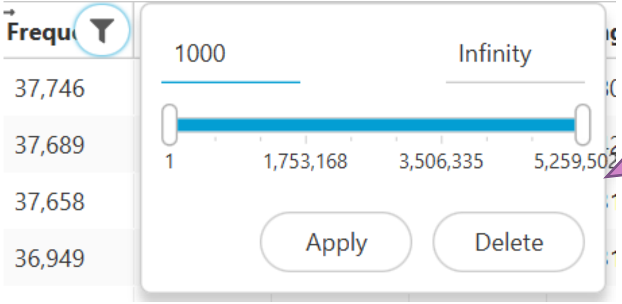
3. The information you can find about the most frequent word_Absolute Freq., Relative Freq., Dispersion_

4. Number of words with a frequency ≥1000 _______7,333________

5. Number of words with a frequency ≥100 _______32,672_______

6. Number of words with a frequency ≥10 _______120,571_______

7. Number of words with a frequency of 1 (so called *hapax legomena*) ________488,054________



Tip: Click on the filter icon and apply an appropriate frequency range to the 'Frequency' column. Click 'Apply'.

Words: Terminology

There are different concepts of a word. In corpus linguistics, terms such as *token, type, lemma* or *lexeme* are often used to denote different senses in which the general term 'word' is used.

Token (running word) is a single occurrence of a word form in a text or corpus.

Type is a unique word form in a corpus.

Lemma denotes all inflected forms belonging to one stem and one word class; in #LancsBox by default, a combination of a headword and a grammatical category (e.g. go + VERB). For example, a lemma 'go' includes the following word forms (types): 'go', 'goes', 'went', 'going' and 'gone'.

Lexeme is a lemma with a particular meaning attached to it, which is necessary to distinguish polysemous words (words with multiple meanings).

Task 2 Understanding dispersion

In this task, you will explore the information about the distribution of words in texts. This is what is called ‘dispersion’ in corpus linguistics. You will be using BNC2014 (whole corpus).

2a What words do you think appear in most English texts? Give some examples.

____

____

2b Based on your intuition, try to guess approximately the percentage of texts the following words occur in:

research _____%   cake _____%   think_____%   hitherto_____% February _____%

Now check your answers using corpus data:

2c In Words, sort the table according to ‘Range %’ and note down top five words and the percentage of texts they appear in.

The (99.15) , to (97.05) , a (96.78) , and (96.57) , of (95.85)

____

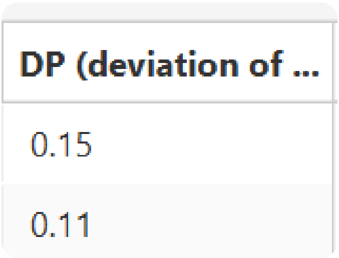
2d Now search for the following words and note down the percentage of texts they occur in:

research 8.86%   cake _1.74%   think_22.65%   hitherto_0.31%   february _5.11%

2e Now search for the following words and note down their DP values:

research DP_0.77_ cake DP__0.88_   think DP_0.52_ hitherto DP_0.98_ february DP_0.88__

Tip: DP stands for ‘Deviation of proportions’. It operates on a scale 0 – 1 with 0 being the most equaly distributed an 1 being extremely unequally distributed.



DP (deviation of ...

0.15

0.11

Task 3

Understanding units: class, pos, sem, lemma

In this task, you will explore different units that a wordlist can be based on. By default, #LancsBox X shows you lowercase types ‘word (lowercase)’. To obtain even more information, the units can be modified.

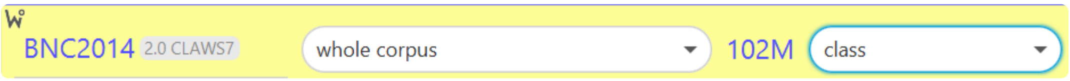
.

3a

In turn, change to unit to ‘class’, ‘pos’ and ‘sem’ and fill in the table below to identify top 5 wordclasses, pos-tags and semantic tags.

Top 5 word classes	Top 5 pos tags	Top 5 semantic tags
SUBST	NN1	Z5
ADJ	JJ	Z8
VERB	II	Z99
PREP	AT	A3
ART	NN2	Z1

Tip: In Words, change the unit in the top bar next to the corpus information.
For the interpretation of pos tags go to: https://ucrel.lancs.ac.uk/claws7tags.html
For the interpretation of semantic tags or to: https://ucrel.lancs.ac.uk/usas/USASSemanticTagset.pdf


BNC2014 2.0 CLAWS7 | whole corpus | 102M | class

3b

Now, change the unit to lemma. In turn, apply filters to the ‘Term’ column to identify most frequent nouns _N, verbs V_, and adjectives _J.

Top 5 nouns _N	Top 5 verbs _V	Top 5 adjectives _J
year	be	good
time	have	new
people	do	other
thing	say	great
way	will	high

3c

Interpret the findings from 3b ‘Top 5 nouns’. What are the concepts that are most frequently discussed in current British English?

Time, people and general concepts – these are typicaly useful across a range of contexts.

It is interesting to see that similar terms - certainly from these areas - will occur both in speech and writing, reflecting the need to frame discourse in terns of time, space, relevant actors (people; down the list: man, woman), and general objects and processes (thing, way).

Task 4 Understanding keywords

Keywords are words that occur with a considerably higher frequency in a given (sub)corpus compared to a reference (sub)corpus. In this task, you will explore keywords used in SMS messages compared to all of elanguage.

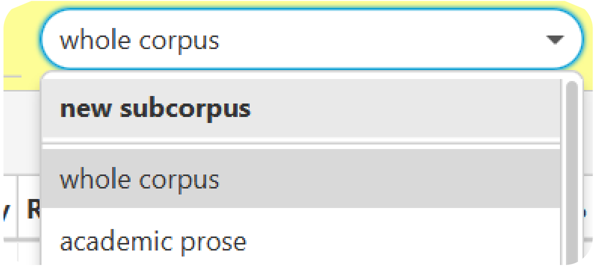
1. First create a SMS messages subcorpus of the BNC2014 and note down its size: ___228k_____________

2. Click on the keyword icon 🔑 and select **BNC2014 elanguage** as a reference corpus.

3. Note down the top 10 keywords, i.e. words typical of SMS messages ____u, lol, haha, yeah, yep, ok, yeh, tho, av, oh _________________

4. Note especially the 1st keyword _________u__________ and its meaning ____2nd person pronoun 'you'

5. Note also the relative frequency of the 1st keyword in SMS messages _______4727.86_____ compared to its frequency in the reference corpus (whole elanguage) ___302.56________________.



Tip: To create a subcorpus, go to the subcorpus manu and click on 'new subcorpus'. Then select the appropriate category (e.g. genre>elanguage, subgenre> SMS messages)

Developed at Lancaster University