



## 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 \_\_\_\_\_
2. The most frequent word in the wordlist is \_\_\_\_\_
3. The information you can find about the most frequent word \_\_\_\_\_
4. Number of words with a frequency  $\geq 1000$  \_\_\_\_\_
5. Number of words with a frequency  $\geq 100$  \_\_\_\_\_
6. Number of words with a frequency  $\geq 10$  \_\_\_\_\_
7. Number of words with a frequency of 1 (so called *hapax legomena*) \_\_\_\_\_

The screenshot shows the 'Frequency' filter interface in the #LancsBox X Words tool. A table with a 'Frequency' column is visible on the left, with values 37,746, 37,689, 37,658, and 36,949. The filter range is set from 1000 to Infinity. The range is visualized as a blue bar on a scale with markers at 1, 1,753,168, 3,506,335, and 5,259,505. Below the scale are 'Apply' and 'Delete' buttons. A filter icon (a funnel) is highlighted in a blue circle next to the 'Frequency' column header.

**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.

---

---

2d

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

research \_\_\_\_\_%    cake \_\_\_\_\_%    think \_\_\_\_\_%    hitherto \_\_\_\_\_%    february \_\_\_\_\_%

2e

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

research DP \_\_\_\_\_    cake DP \_\_\_\_\_    think DP \_\_\_\_\_    hitherto DP \_\_\_\_\_    february DP \_\_\_\_\_

**Tip:** DP stands for 'Deviation of proportions'. It operates on a scale 0 – 1 with 0 being the most equally distributed and 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

**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 go to: <https://ucrel.lancs.ac.uk/usas/USASSemanticTagset.pdf>



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 <u>_N</u>	Top 5 verbs <u>_V</u>	Top 5 adjectives <u>_J</u>

3c

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

---

---


---

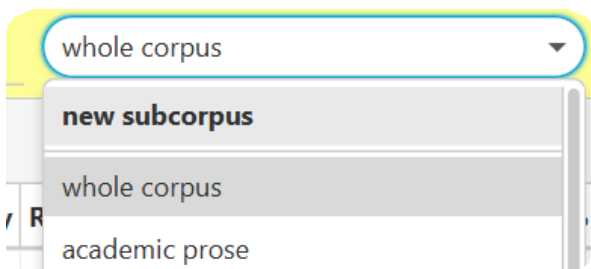
---

## 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: \_\_\_\_\_
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 \_\_\_\_\_  
\_\_\_\_\_
4. Note especially the 1<sup>st</sup> keyword \_\_\_\_\_ and its meaning \_\_\_\_\_
5. Note also the relative frequency of the 1<sup>st</sup> keyword in SMS messages \_\_\_\_\_ compared to its frequency in the reference corpus (whole elanguage) \_\_\_\_\_.



**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

