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_________

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).


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 equally distributed and 1 being extremely unequally distributed.
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.

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.

<table>
<thead>
<tr>
<th>Top 5 word classes</th>
<th>Top 5 pos tags</th>
<th>Top 5 semantic tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>SUBST</td>
<td>NN1</td>
<td>Z5</td>
</tr>
<tr>
<td>ADJ</td>
<td>JJ</td>
<td>Z8</td>
</tr>
<tr>
<td>VERB</td>
<td>JJ</td>
<td>Z99</td>
</tr>
<tr>
<td>PREP</td>
<td>AT</td>
<td>A3</td>
</tr>
<tr>
<td>ART</td>
<td>NN2</td>
<td>Z1</td>
</tr>
</tbody>
</table>

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/claws7/tags.html
For the interpretation of semantic tags or to: https://ucrel.lancs.ac.uk/usas/USASSemanticTagset.pdf

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_.

<table>
<thead>
<tr>
<th>Top 5 nouns <em>N</em></th>
<th>Top 5 verbs <em>V</em></th>
<th>Top 5 adjectives <em>J</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>year</td>
<td>be</td>
<td>good</td>
</tr>
<tr>
<td>time</td>
<td>have</td>
<td>new</td>
</tr>
<tr>
<td>people</td>
<td>do</td>
<td>other</td>
</tr>
<tr>
<td>thing</td>
<td>say</td>
<td>great</td>
</tr>
<tr>
<td>way</td>
<td>will</td>
<td>high</td>
</tr>
</tbody>
</table>

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 typically 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 terms of time, space, relevant actors (people; down the list: man, woman), and general objects and processes (thing, way).
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)