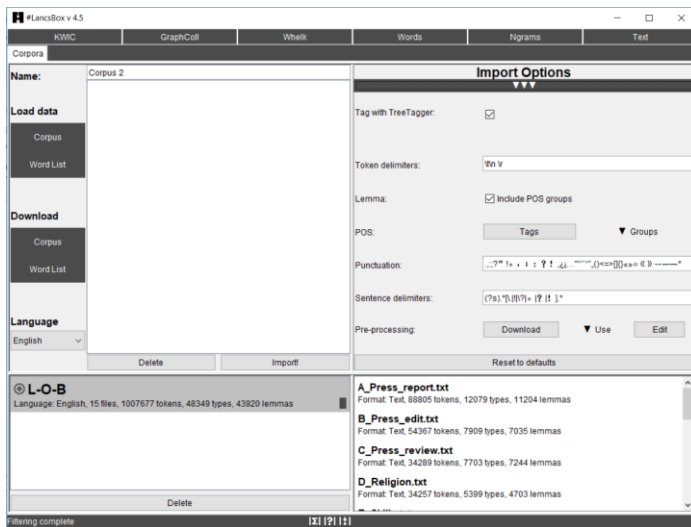


## 2 Loading and importing data

Data can be loaded and imported into #LancsBox on the 'Corpora' tab. This tab opens automatically when you run #LancsBox. #LancsBox works with corpora in different formats (.txt, .xml, .doc, .docx, .pdf, .odt, .xls, .xlsx, .zip etc.) and with wordlists (.csv). There are two options for loading corpora and wordlists: i) load (your own) data and ii) download corpora and wordlists that are distributed with #LancsBox.

### 2.1 Visual summary of Corpora tab



**Top panel:** Importing corpora and wordlists

**You can:**

- Select your corpus or wordlist to load.
- Download a corpora and wordlists distributed with #LancsBox.
- Select language.
- Review POS tags.
- Review punctuation marks and sentence delimiters.
- Set-up pre-processing via customisable scripts.

**Bottom panel:** Working with corpora and wordlists

**You can:**

- Activate or delete imported corpora or wordlists.
- Review corpus and text size (tokens, types, lemmas).
- Preview texts.
- Save processed corpora with pos-tags etc.

### 2.2 Load your corpora and wordlists

#LancsBox allows you to work easily with your own corpora and wordlists. These corpora are those stored on your computer or at a location accessible from your computer (memory stick, shared drive, dropbox, cloud etc.).

1. In the Corpora tab, left-click on 'Corpus' or 'Word List' under 'Load data', depending on whether you want to load a corpus or a wordlist.
2. This will open a window where you can navigate to the location (folder) where your corpus or wordlist is stored.
3. You can select a specific file, select multiple files by holding down Ctrl and left-clicking on your chosen files, or select all files in the folder by holding down Ctrl + A.
4. Left-click 'Open' to load your files.
5. Select the language of your corpus or wordlist. #LancsBox supports automatic lemmatisation and POS tagging in multiple languages. This is done using Tree Tagger. If your language is not listed, select 'Other'; in this case, automatic lemmatisation and POS tagging will be disabled.
6. [Optional: You can review/change the import options by left-clicking on a bar with three triangles (▲▲▲). In most cases, you can use the default options.]

7. Left-click 'Import!' to import your corpus into #LancsBox. By default, #LancsBox automatically adds POS tags to the corpus.

## 2.3 Supported file formats

---

#LancsBox supports different file formats (.txt, .xml, .doc, .docx, .pdf, .odt, .xls, .xlsx, .zip and many others) of corpus files. #LancsBox automatically extracts and processes text available in corpus files. For wordlists, #LancsBox assumes the comma-delimited file format (.csv).

---

1. Corpus formats: .txt, .xml, .doc, .docx, .pdf, .odt, .xls, .xlsx, .zip – full list: [Apache Tika](#).
2. Wordlist format: csv (see example below).

```
Corpus: BNC| Language: English| 4055 files| 96996843 tokens| 662414 types| 716618 lemmas|
>Type", "Frequency: 01 - Freq", "Dispersion: 01_CV"
"the", "6054524.000000", "0.286889"
"of", "3049295.000000", "0.400166"
"and", "2622080.000000", "0.263099"
"to", "2599355.000000", "0.223254"
"a", "2168976.000000", "0.221813"
"in", "1945319.000000", "0.333547"
```

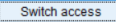
## 2.4 Download #LancsBox corpora and wordlists

---

#LancsBox allows you to work with existing corpora that are freely distributed with #LancsBox under a specific license. Two modes for corpus sharing are available: i) open access and ii) restricted access. We are constantly adding more corpora to this list.

---

1. In the corpora tab, left-click on 'Corpus' or 'Word List' under 'Download'.
2. This will open a window where you can select corpora or wordlists distributed with #LancsBox. By left-clicking on a corpus, you will be shown additional information about the corpus or wordlist, including the language, date, text type, license etc.
3. Review and agree with the corpus license.
4. Left-click 'Download' to download the selected corpus or wordlist.
5. Left-click 'Import!' to import your corpus into #LancsBox. By default, #LancsBox automatically adds POS tags to the corpus.

► **Note:** To switch between open and restricted access corpora, use the 'Switch access' button in the bottom left corner (  ). Restricted access corpora are distributed as encrypted and have several display and usage restrictions. For example, they cannot be displayed in the Text tool or saved to the local computer.

## 2.5 Working with corpora and wordlists

---

All corpora and wordlists that have been imported into #LancsBox are displayed in the bottom panel on the 'Corpora' tab. This panel allows reviewing corpora, previewing files and fast reloading of corpora and wordlists when #LancsBox is closed and re-opened.

---

1. If you have imported a corpus (📁) or wordlist (📄) it will appear in the bottom panel, alongside any other corpora or wordlist you have already imported. These can be removed by left-clicking 'delete'. In the bottom-right section, you can view the corpus structure: the individual text files that the corpus is composed of.
2. In the bottom panel (bottom left window), the default corpus can also be specified. The default corpus is a corpus that #LancsBox offers as a default choice in the individual modules. The default corpus can be specified by left-double-clicking on the name of the corpus; a filled rectangle (■) will appear next to the name of the default corpus.
3. If #LancsBox is closed, the corpora and wordlists will remain imported but will be unloaded. To activate (reload) the corpora or wordlists for use, left-double-click on the corpora or wordlists.
4. You can also preview the files by right-clicking on them. They will appear in the Text tool (see Section 8). The list of files (including the info about their size) can also be copied (Ctrl/Command+C) and pasted (Ctrl/Command+V) into a spreadsheet or text document.
5. Corpora are now ready to be analysed using five modules: KWIC, Whelk, GraphColl, Words and Text. Wordlists can be used in the Words tool.

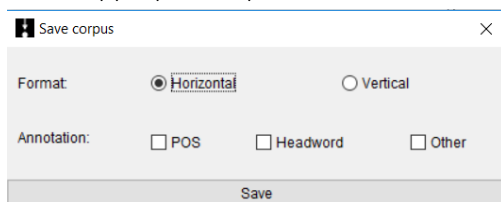
## 2.6 Saving corpora

---

#LancsBox saves corpora in the horizontal or the vertical format.

---

1. Right-click on the corpus which you wish to save.
2. Select appropriate options.



3. Click 'Save'.

## 2.7 Pre-processing of corpora (Advanced users).

---

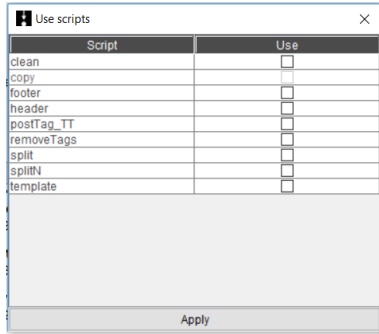
#LancsBox allows pre-processing data as part of the import procedure. This is set up in the 'Import options' under 'Pre-processing'. Data can be modified in different ways using a variety of Groovy scripts, which are fully customisable.

---

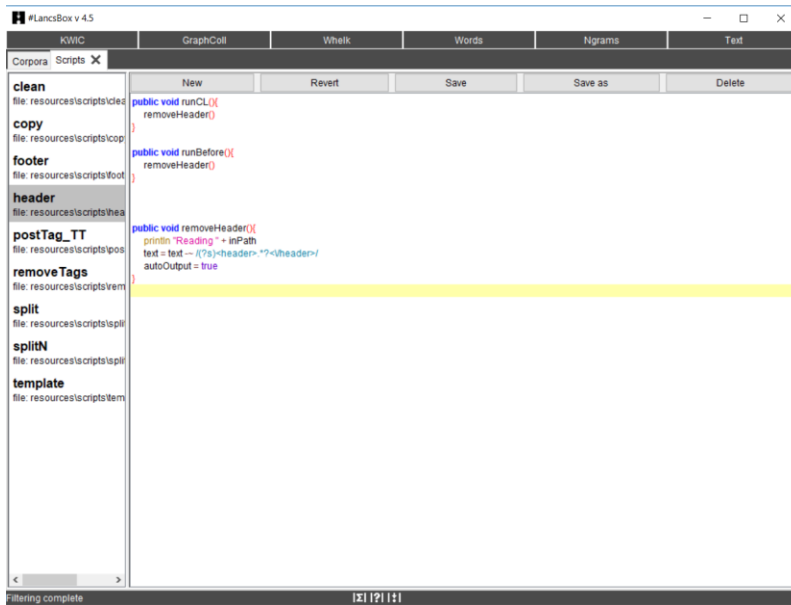
1. Under 'Pre-processing' three options are available:



2. 'Download' allows the user to download scripts and their newest versions available from the #LancsBox website.
3. 'Use' displays a list of currently available scripts and a checkbox next to each script for the user to indicate, which scripts will be used in the pre-processing stage.



4. 'Edit' displays the scripts in a script editor, which allows modifying existing scripts and creating new scripts.



5. The structure of a script is as follows. More information about the Groovy scripting language can be found at <http://groovy-lang.org>.

Script	Comments
<pre>public void runCL(){     println "Ran on the command line." }</pre>	Scripts run via the command line.
<pre>public void runBefore(){     println "Ran as a pre-process script." }</pre>	Scripts run when the files are being loaded. This allows splitting files, deleting or changing texts or structuring elements e.g. xml tags.

<pre>public void runAfter(Token token){     println "Ran after the tagging step." }  public void removeHeader(){     println "Reading " + inPath     text = text ~/((?s)&lt;header&gt;.*?&lt;/header&gt;/     autoOutput = true }</pre>	<p>Scripts run after part-of-speech tagging. This allows modifying the output of the Tree Tagger, e.g. correcting tagging errors.</p> <p>An example of a simple script deleting header indicated by &lt;header&gt;&lt;/header&gt; tags in text.</p>
---	---

► **Did you know?**

The Brown corpus and the LOB (Lancaster-Oslo/Bergen) corpus are one of the first modern corpora stored and processed on computers. Each consists of one million running words (tokens), a size that was very ambitious at the time of their compilation. Brown was compiled in the 1960s by Henry Kučera and W. Nelson Francis at Brown University (US). It was originally stored and processed on IBM punch cards. In the early 1970s, a British counterpart to the Brown corpus was compiled as a collaboration between Lancaster University (UK) and two Norwegian universities: Oslo and Bergen. The project was initiated by Geoffrey Leech from Lancaster University.