Manual for the Web Corpus Compiler (WCC)

Version 1.0

Author:
Martin Weisser

June 2022

Contents

1 Introduction

The Web Corpus Compiler (WCC for short), as its name implies, is a tool for creating corpora from the web. Unlike some tools that automatically crawl the web and compile corpora based on specific seed terms, and where you have no initial choice over what gets collected, all decisions on the composition of the corpus in the WCC are made by first inspecting suitable pages suggested by one of three possible search engines, adding these to a list of files to download in a simple and convenient way, and then downloading the relevant URLs, with all successful downloads being logged to a spreadsheet. Thus downloaded HTML files can then be converted to UTF-8-encoded plain text files, whereby a fair degree of linguistically motivated boilerplate removal is carried out, and the results can be further edited within the WCC by comparing them side-by-side to the original HTML files. Although this process is obviously more time-consuming than fully automated crawling and compilation, the ability to choose your data carefully should arguably guarantee a higher quality for the resulting corpus.

The design of the WCC is essentially based on one of my earlier tools, ICEweb (v. 2; written in Perl/Tk), which was predominantly intended for compiling new ICE components from the web, as well as carrying out some basic analyses on these. As ICEweb now – for some unknown reason – no longer appears to be running on some updated versions of Windows, most notably Chinese ones, I decided to write a more general replacement for it in Python and PyQt, with the added advantage of the WCC now also being useable on Macs, as well as potentially on Linux. In addition, all searching and viewing of candidate pages can now be carried out from within the tool itself, without ever needing to start an external browser, as can be seen in Figure 1, which depicts the WCC after startup, with Google being selected as a search engine.

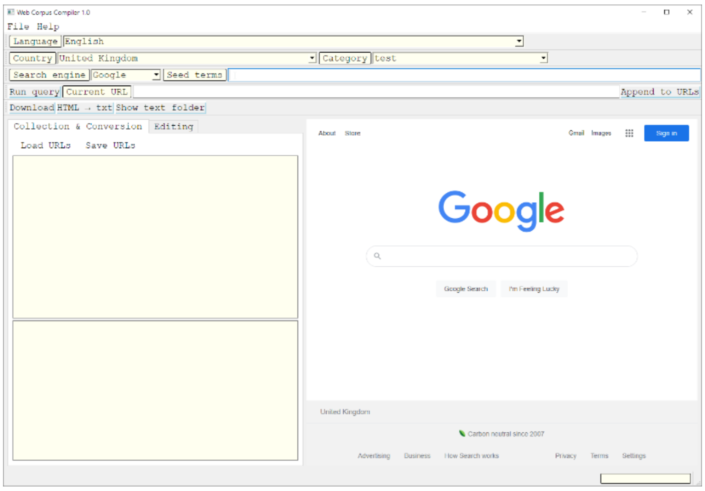

Language English
Country United Kingdom
Category test
Search engine Google

Figure 1 the WCC after startup

The WCC interface is essentially separated into three parts, the toolbar for adjusting relevant settings and triggering various actions, a notebook that contains tabs for ‘Collection & Conversion’ activities and for ‘Editing’ the corpus data on the left-hand side, and a browser window on the right-hand side. In addition, there’s a status bar at the bottom. The ‘Collection & Conversion’ tab depicted in Figure 1 is sub-divided into an editor window where messages pertaining to download or conversion processes will appear, and the URL editor used for compiling a list of URLs to be included in a download. The layout for the ‘Editing’ tab will be described in the relevant section further down.

2 Steps in Compiling a Corpus

In order to compile a corpus for a particular language, country, and category, you only need to follow a few simple steps detailed in the following sections.

2.1 Selecting a Language, Country & Category

The first step in searching for suitable data and compiling your corpus consists in selecting the language for the web pages you want to find, the country domain you want to search in, and the text category. Unless you work with multiple languages and domains, the first two may only need to be configured once, while the third one is likely to change more frequently, but can also be added to the settings as a default, as described in section 3 below. Choosing a language from the ‘Language’ dropdown list is depicted in Figure 2.

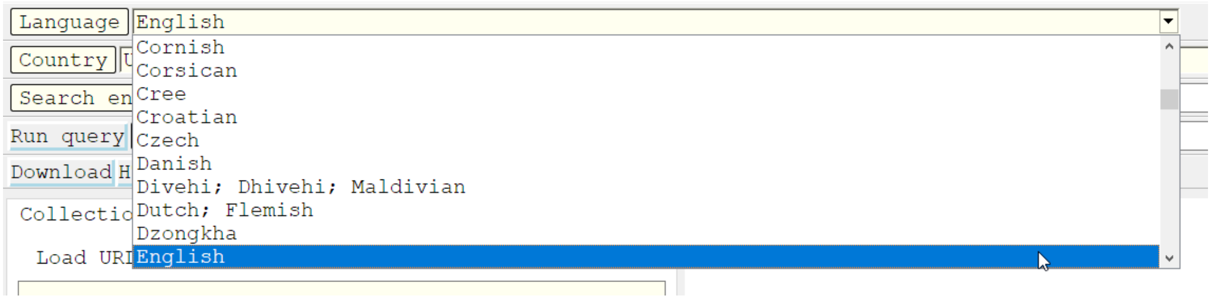


Figure 2 selecting a language

Selecting a country is done via the ‘Country’ dropdown list, shown in Figure 3.

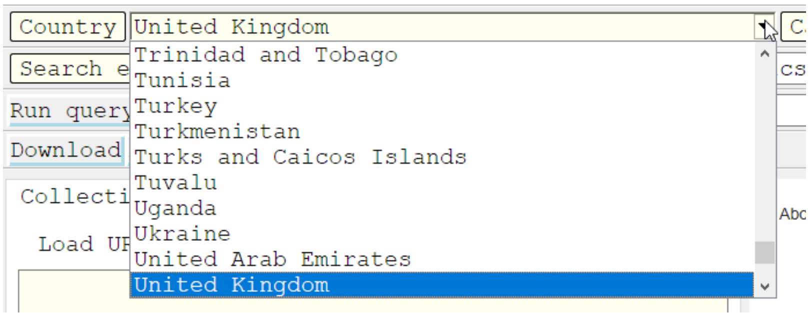


Figure 3 selecting a country

Equally, choosing a category is done directly via the 'Category' dropdown, at least if you're choosing a pre-existing one. If the category you want to use doesn't exist yet, you first need to create it using the configurations dialogue as, again, described in section 3 below.

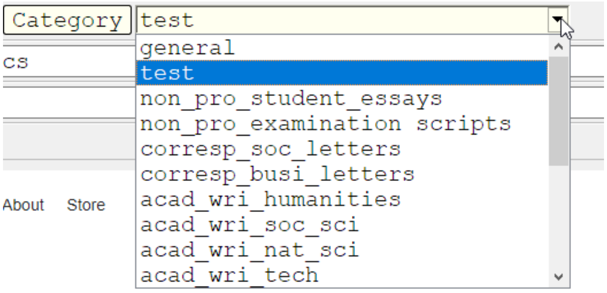


Figure 4 selecting a category

Once these three options have been chosen, the first two can initially be used to target the search towards the particular language and domain in the search engine. All three of these, though, are used to automatically create the relevant resources, such as file paths, etc., for downloading and storing the data.

2.2 *Selecting a Search Engine & Adding Seed Terms*

To begin the process of identifying relevant web pages, you next need to select your favourite search engine out of the options provided. These currently comprise Google, DuckDuckGo, and Bing, as can be seen in Figure 5.

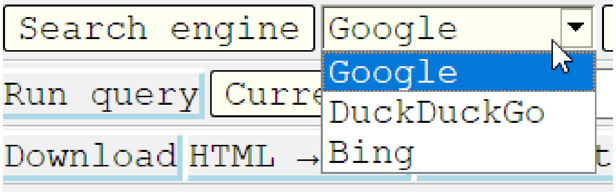
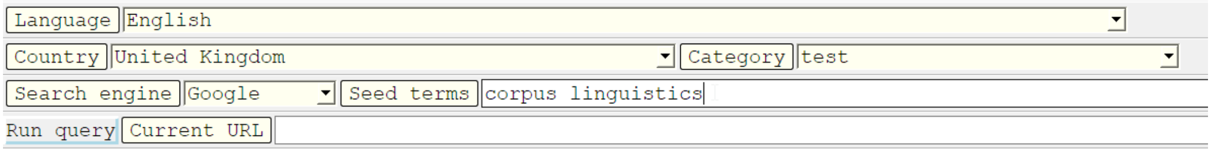


Figure 5 selecting a search engine.

As for the other options, the default can again be set via the WCC configuration. My personal recommendation here would be Google, with a reasonable second being DuckDuckGo, as they tend to provide far better results. Bing has essentially been added here only as a fallback because the former two may not be available in some countries.

2.3 Running a Query

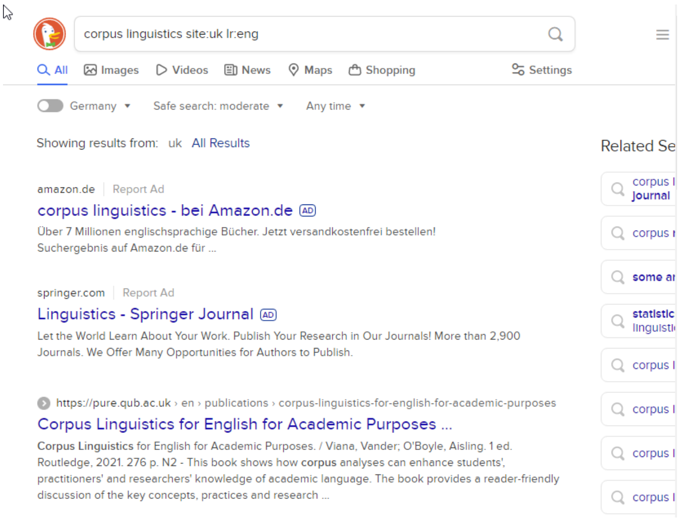
To run a query, you should first add some seed terms, i.e. terms that are likely to occur in documents reflecting your category, as visible in the text box next to ‘Seed terms’ in Figure 6, where I added the terms corpus and linguistics to identify pages related to the subject, as I would do in a normal web search.



Language: English
Country: United Kingdom
Category: test
Search engine: Google
Seed terms: corpus linguistics

Figure 6 specifying seed terms

To run the web search itself, you click ‘Run query’, which will open the selected engine in the browser window, adding the relevant advanced search parameters for the language and domain automatically in the correct format for this engine in addition to the seed terms. For all engines but Google, the search will automatically be run, too, and the results displayed. For Google, this will only fill in the query terms, but you’ll still need to click on the button to start the search. Figure 7 shows the query results for the seed terms *corpus* and *linguistics* with the country setting for the *UK* and language *English*.



corpus linguistics site:uk lr:eng

Showing results from: uk All Results

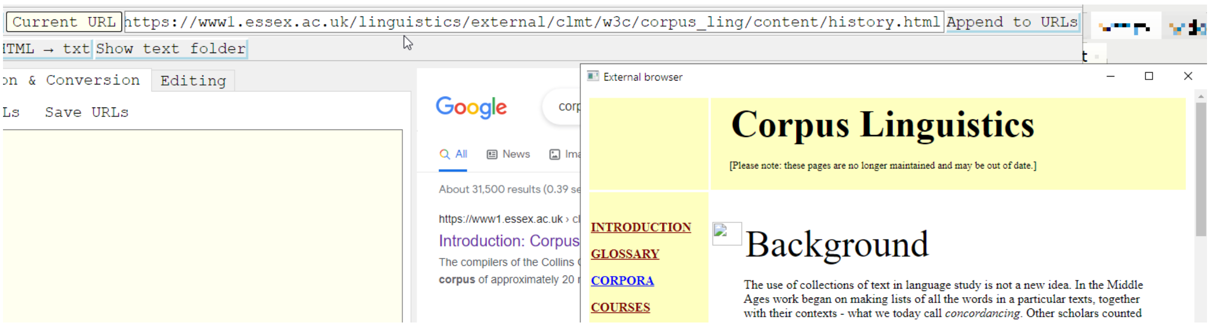
amazon.de | Report Ad
corpus linguistics - bei Amazon.de AD
Über 7 Millionen englischsprachige Bücher. Jetzt versandkostenfrei bestellen! Suchergebnis auf Amazon.de für ...

springer.com | Report Ad
Linguistics - Springer Journal AD
Let the World Learn About Your Work. Publish Your Research in Our Journals! More than 2,900 Journals. We Offer Many Opportunities for Authors to Publish.

https://pure.qub.ac.uk › en › publications › corpus-linguistics-for-english-for-academic-purposes
Corpus Linguistics for English for Academic Purposes ...
Corpus Linguistics for English for Academic Purposes. / Viana, Vander; O'Boyle, Aisling. 1 ed. Routledge, 2021. 276 p. N2 - This book shows how **corpus** analyses can enhance students', practitioners' and researchers' knowledge of academic language. The book provides a reader-friendly discussion of the key concepts, practices and research ...

Figure 7 query results in DuckDuckGo

Clicking on any of the links in the query results browser window will open the relevant link for inspection inside a new floating browser window (Figure 8), as well as copy its URL to the 'Current URL' textbox (Figure 9), from where it can be appended to the URL list simply by clicking the 'Append to URLs' button if it fits your criteria for inclusion in the corpus.



Current URL https://www1.essex.ac.uk/linguistics/external/clmt/w3c/corpus_ling/content/history.html

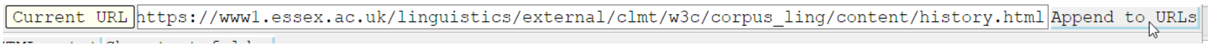
Corpus Linguistics

[Please note: these pages are no longer maintained and may be out of date.]

Background

The use of collections of text in language study is not a new idea. In the Middle Ages work began on making lists of all the words in a particular texts, together with their contexts - what we today call *concordancing*. Other scholars counted

Figure 8 link opened in new floating browser window



Current URL https://www1.essex.ac.uk/linguistics/external/clmt/w3c/corpus_ling/content/history.html

Figure 9 URL copied to Current URL textbox, ready for appending to URL list

Note that the floating browser window will sometimes get hidden if you perform other operations on your computer, so that you may need to find it again if it isn't immediately visible after clicking on a link.

2.4 Handling URLs

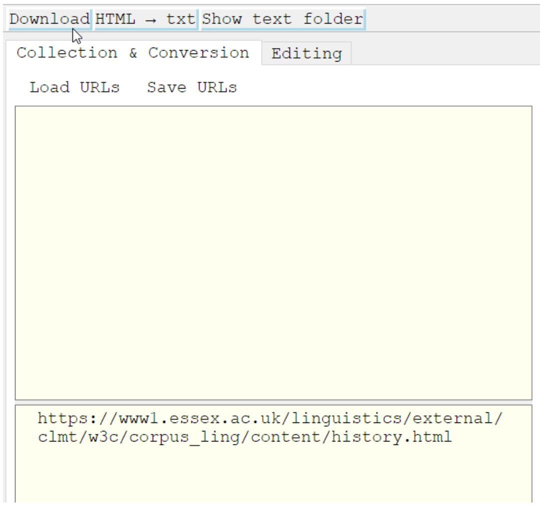
Once you've compiled a list of links in the URL editor window, this can immediately be used to trigger a download, but in most cases, you'll probably want to save the file first by clicking on the 'Save URLs' button. Doing so will automatically add the contents of the URL editor to the file called 'URLs.txt' in the relevant sub-folder for the chosen language, country, and category in the 'downloads' folder within the WCC program folder.

The file, as well as any relevant file paths, will automatically be created, based on your active selections. If the file already exists, but you haven't loaded it via the 'Load URLs' button prior to adding URLs to the editor window, the WCC will also automatically open the file and append the URLs from the editor window to ensure that you don't accidentally lose any URLs collected in a prior compilation session.

As normally compiling a web corpus will consist of a number of sessions, it generally makes sense to load an existing URL file at the beginning of each new session, prior to downloading any data. As in creating the file, the WCC will automatically identify the correct file to load, based on your active selection, so you don’t need to navigate any folders.

2.5 Downloading files

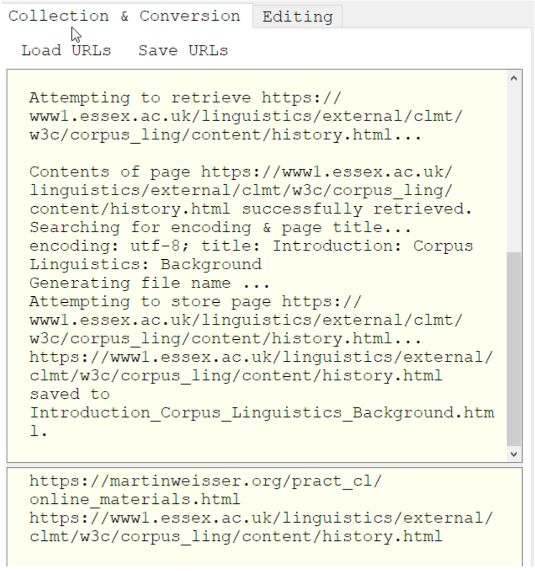
To download one or more files, all you need to do is have a few URLs listed in the URL editor window, and then click the ‘Download’ button immediately above the tab, as depicted in Figure 10.



```
https://www1.essex.ac.uk/linguistics/external/
clmt/w3c/corpus_ling/content/history.html
```

Figure 10 starting a download

This will trigger a download process for all URLs listed in the editor window, line-by-line, reporting the progress for each file step-by-step until all URLs have been processed.



```
Attempting to retrieve https://
www1.essex.ac.uk/linguistics/external/clmt/
w3c/corpus_ling/content/history.html...

Contents of page https://www1.essex.ac.uk/
linguistics/external/clmt/w3c/corpus_ling/
content/history.html successfully retrieved.
Searching for encoding & page title...
encoding: utf-8; title: Introduction: Corpus
Linguistics: Background
Generating file name ...
Attempting to store page https://
www1.essex.ac.uk/linguistics/external/clmt/
w3c/corpus_ling/content/history.html...
https://www1.essex.ac.uk/linguistics/external/
clmt/w3c/corpus_ling/content/history.html
saved to
Introduction_Corpus_Linguistics_Background.htm
l.
```

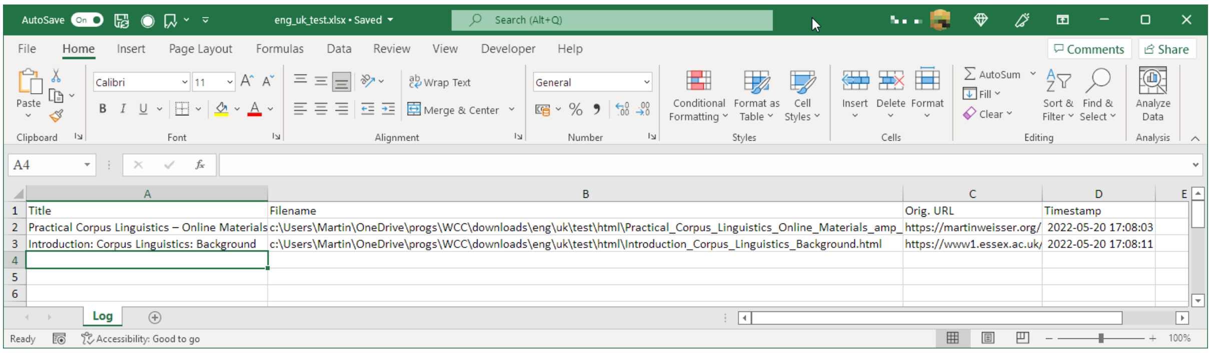
```
https://martinweisser.org/pract_cl/
online_materials.html
https://www1.essex.ac.uk/linguistics/external/
clmt/w3c/corpus_ling/content/history.html
```

Figure 11 download report progress

During this process, the WCC will access the relevant URL, attempt to retrieve it from the web server, identify its encoding, and then store the HTML file, encoded as UTF-8, in the ‘html’ sub-folder of the category folder.

If, while reading in the HTML file, an unknown error occurs, the download will be aborted to prevent files that have been encoded incorrectly from being stored. If the file can be read successfully, though, as should normally be the case, then the WCC will create a suitable file name from either the title tag of the page or the name of the HTML file in case there isn’t a title, which is rare these days,

For successfully downloaded files, the title, file name, and the original URL will be recorded in a spreadsheet inside the category folder, together with a timestamp for the download date and time.



	A	B	C	D
1	Title	Filename	Orig. URL	Timestamp
2	Practical Corpus Linguistics – Online Materials	c:\Users\Martin\OneDrive\progs\WCC\downloads\eng\uk\test\html\Practical_Corpus_Linguistics_Online_Materials_amp_	https://martinweisser.org/	2022-05-20 17:08:03
3	Introduction: Corpus Linguistics: Background	c:\Users\Martin\OneDrive\progs\WCC\downloads\eng\uk\test\html\Introduction_Corpus_Linguistics_Background.html	https://www1.essex.ac.uk/	2022-05-20 17:08:11

Figure 12 download log in spreadsheet

Additional files can always be downloaded and added to the corpus later by adding URLs to the URL file. If files have already been downloaded, you’ll get the option of overwriting or skipping them through a dialogue, but overwriting really only makes sense in case you assume that there is a newer, updated, version. For any additional files, the information will also be appended to the spreadsheet, but if you download a newer version, you’ll need to remove the row containing the information for the previous download manually. The file names for the spreadsheets always consist of abbreviated language and country ids and the category name, joined by underscores, so they should be easy to find.

2.6 Converting HTML to text

Once you’ve downloaded a suitable number of HTML files, you can convert them to plain text files by clicking on the ‘HTML → txt’ button as shown in Figure 13.

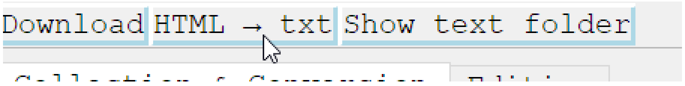
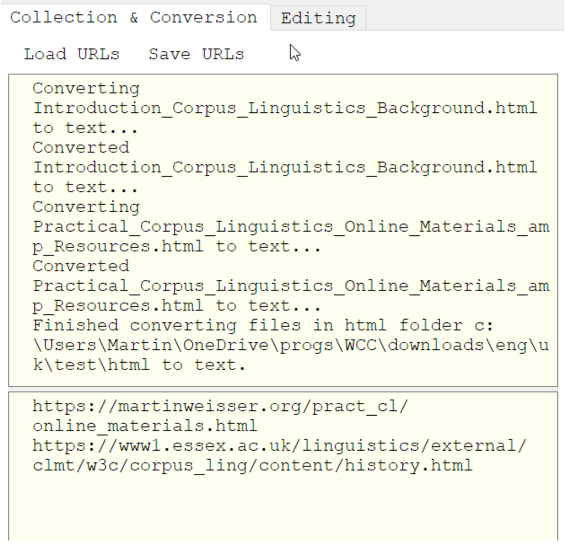


Figure 13 converting from HTML to text

This will automatically convert all files located in the ‘html’ sub-folder and output them to the ‘txt’ folder. The conversion process not only strips away the HTML tags, but also attempts to remove any boilerplate content, i.e. navigational links, scripts, etc., using purely linguistic heuristics. The conversion process will also be logged in the log window, as can be seen in Figure 14.



```
Converting
Introduction_Corpus_Linguistics_Background.html
to text...
Converted
Introduction_Corpus_Linguistics_Background.html
to text...
Converting
Practical_Corpus_Linguistics_Online_Materials_am
p_Resources.html to text...
Converted
Practical_Corpus_Linguistics_Online_Materials_am
p_Resources.html to text...
Finished converting files in html folder c:
\Users\Martin\OneDrive\progs\WCC\downloads\eng\u
k\test\html to text.
```

```
https://martinweisser.org/pract_cl/
online_materials.html
https://www1.essex.ac.uk/linguistics/external/
clmt/w3c/corpus_ling/content/history.html
```

Figure 14 conversion logging

This process may already leave you with reasonably cleaned-up text files for your corpus, but in order to enhance the quality of your corpus data further, the WCC also provides convenient ways for editing the text files by comparing them side-by-side to the original web pages as described in the next section.

2.7 Editing corpus data further

To begin editing your corpus data, you click on the ‘Show text folder’ button (Figure 15), which will automatically switch to the Editing tab – if this isn’t already active –, and list all files in the ‘text’ folder for your category, provided of course it’s not empty.

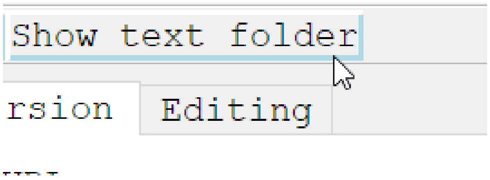


Figure 15 opening the text folder

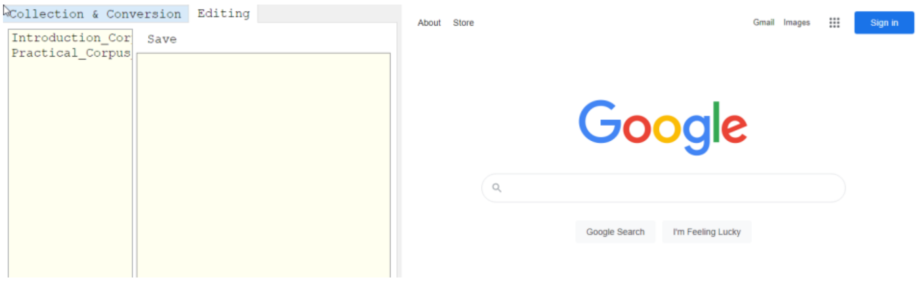
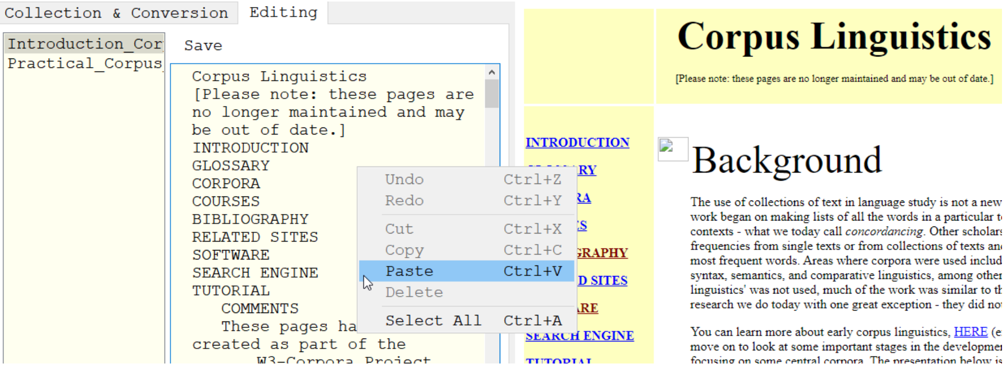


Figure 16 the editing tab layout

Just like the ‘Collection & Conversion’ layout, the ‘Editing’ tab layout is again sub-divided into two sub-panes to the left of the browser pane. Here, the left-hand pane contains the listing of the files in the ‘txt’ folder, while the right-hand pane houses an editor that allows you to post-process the converted text files further.

If you double-click any of the filenames in the text listing on the left, the corresponding text file will be opened in the editor pane, while its HTML counterpart will be displayed in the browser pane.



```
Corpus Linguistics
[Please note: these pages are
no longer maintained and may
be out of date.]
INTRODUCTION
GLOSSARY
CORPORA
COURSES
BIBLIOGRAPHY
RELATED SITES
SOFTWARE
SEARCH ENGINE
TUTORIAL
    COMMENTS
    These pages ha
created as part of the
         W3-Corpora Project
```

Corpus Linguistics

[Please note: these pages are no longer maintained and may be out of date.]

Background

The use of collections of text in language study is not a new work began on making lists of all the words in a particular t contexts - what we today call *concordancing*. Other scholars frequencies from single texts or from collections of texts an most frequent words. Areas where corpora were used includ syntax, semantics, and comparative linguistics, among other linguistics' was not used, much of the work was similar to th research we do today with one great exception - they did no

You can learn more about early corpus linguistics, HERE (e move on to look at some important stages in the developmer focusing on some central corpora. The presentation below is

Figure 17 post-processing converted text

Viewing the two files side-by-side makes it easier to see if any of the remaining text should either be deleted as irrelevant or, conversely, any text that may have been deleted as part of the boilerplate removal might need to be added. In the above example, for instance the note in square brackets could be removed as it doesn’t really reflect the topic of corpus linguistics, but in a sense is a from of meta information related to the page content. Similarly, the navigational links in all-caps were not identified by the boilerplate removal algorithm due to the section containing them not having been marked as a navigational part of the page, as the page uses a rather old-fashioned tabular layout where the links appear inside separate paragraphs and not even as part of a list structure, and even after removing some formatting tags, the otherwise excellent *BeautifulSoup* module I use for manipulating and extracting from the pages couldn't identify the instances of paragraphs containing only links in order to remove them, at least not in this particular document. In future versions, I will try to fine-tune that part of the conversion process, though, to achieve more accuracy in the boilerplate removal process.

As Figure 17 also shows, the editor pane contains a fully-functional plain text editor that allows you to carry out copy-and-paste operations, and supports undo and redo. These need to be accessed through the context menu, though, i.e. by using a right mouse click. The only button that is currently defined for the editor is the 'Save' button, which, as its name suggests, allows you to save any changes you make to the document.

3 Customising the WCC

To customise the WCC, you click on the 'Configure' menu item (Figure 18), which will open the configuration dialogue shown in Figure 19.

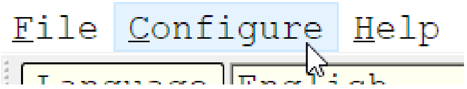
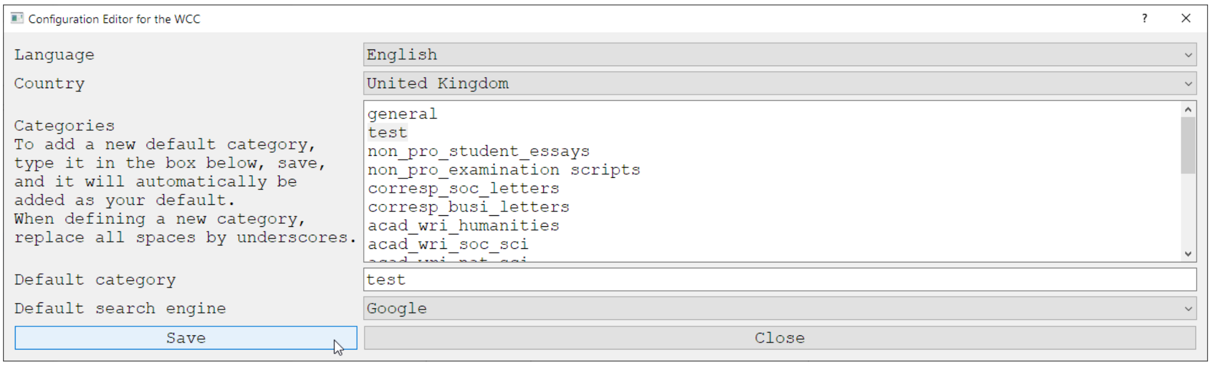


Figure 18 the 'Configure' menu item



Language: English

Country: United Kingdom

Categories
To add a new default category,
type it in the box below, save,
and it will automatically be
added as your default.
When defining a new category,
replace all spaces by underscores.

general
test
non_pro_student_essays
non_pro_examination scripts
corresp_soc_letters
corresp_busi_letters
acad_wri_humanities
acad_wri_soc_sci

Default category: test

Default search engine: Google

Figure 19 the configuration dialogue

As can be seen in the above figure, it's possible to configure defaults for the language, country, category, and search engine. In addition, the list of categories can be edited, and new categories added. To add a new default category, all one has to do is type it into the 'Default category' box and click ‘Save’. The new default will then be added to the list of categories and be set for both the current and future sessions. All other options can simply be picked from the predefined lists.

4 List of Keyboard Shortcuts

The following keyboard shortcuts have been defined for convenient data handling.

Shortcut	**Function**
F3	Load URLs
F2	Save URLs
Ctrl + s	Save file in editor
Ctrl+ r	Run query
Ctrl + p	Append to URLs
Ctrl + d	Download
Ctrl + h	Convert HTML → text
Ctrl + t	Show text folder
F1	Help (opens manual)
Ctrl + q	Exit WCC