Corpus Building with TEC Tools

Version 2.2 (May 2011)

Contents **Page**

Notes

- This tutorial has been created to provide informal assistance to users of the TEC Tools Corpus Software.
- This tutorial was created using a Mac operating system. Further explanatory notes will be provided where there are significant differences for Windows users.
- At the time of writing, TEC Tools is capable of working with English, European languages using Latin alphabets, and Japanese. Parsing ability for other languages will be added in the near future.
- Both the software and this tutorial document are under development and you are advised to check on-line regularly to ensure you are using the most up-to-date versions before you start work.
- Since the software is freely available to users it is not possible to offer technical support. However, a user forum is planned for the near future.
- TEC Tools is Open Source software. For more information: http://modnlp.berlios.de/doc/

Disclaimer

- The authors accept no responsibility for any problems that may arise through the use of the TEC Tools software, nor the information provided in this document.

Contact

- If you have any comments about this tutorial document please email sally.marshall@postgrad.manchester.ac.uk

Contributors
Dr. Saturnino Luz (Trinity College Dublin)
Sofia Malmatidou (University of Manchester)
Sally Marshall (University of Manchester)

Editor
Sally Marshall (University of Manchester)

1. Scanning and Converting Images to Text

Once you have selected the texts you plan to include in your corpus, and obtained permission from the copyright holder (if necessary), you will need to convert your texts into a computer-readable, or digital/electronic, format.

If your data is on printed paper, you will need to scan the printed documents and use OCR (Optical Character Recognition) Software to extract text from them. If you are using texts that are already computer-readable (e.g. text taken from the internet, e-Books, searchable PDF files, Word files etc) you can move on to step 2.

1.1 Scanning Documents

Depending on the type and volume of printed material you are working with you may choose to use either a flat-bed scanner or a scanner with a document feeder.

Flat-bed scanner
Place one page at a time onto the glass scanner bed
Most home-office scanners are flat-bed scanners

Scanner with document feeder
Place multiple pages in the document feeder
Many print shops and university departments have scanners with document feeders

If you have a large number of books to scan you may want to consider removing the pages from their binding so you can use a document feeder. Many print shops are able to remove pages from book bindings using a guillotine, and may also be able to rebind books after scanning. When you scan multi-page documents you should save all pages as one PDF file.

Whichever type of scanner you decide to use you should bear in mind that the quality of the scanned images should be as high as possible in order to ensure good results when using OCR software. Therefore you should check that your scanner is capable of producing images of a sufficiently high quality. Some other points to consider include:

Paper quality
Thicker paper is less prone to “show-through” (where text from the other side of the page shows through). For this reason, hard-backed editions of books may produce better results than paperbacks.

Print quality
The better the quality of the printed text, the better the quality of the output. Avoid poor quality print, for example with “bleed”.

Font
Commonly used fonts often produce better results than unusual fonts

Scanner settings

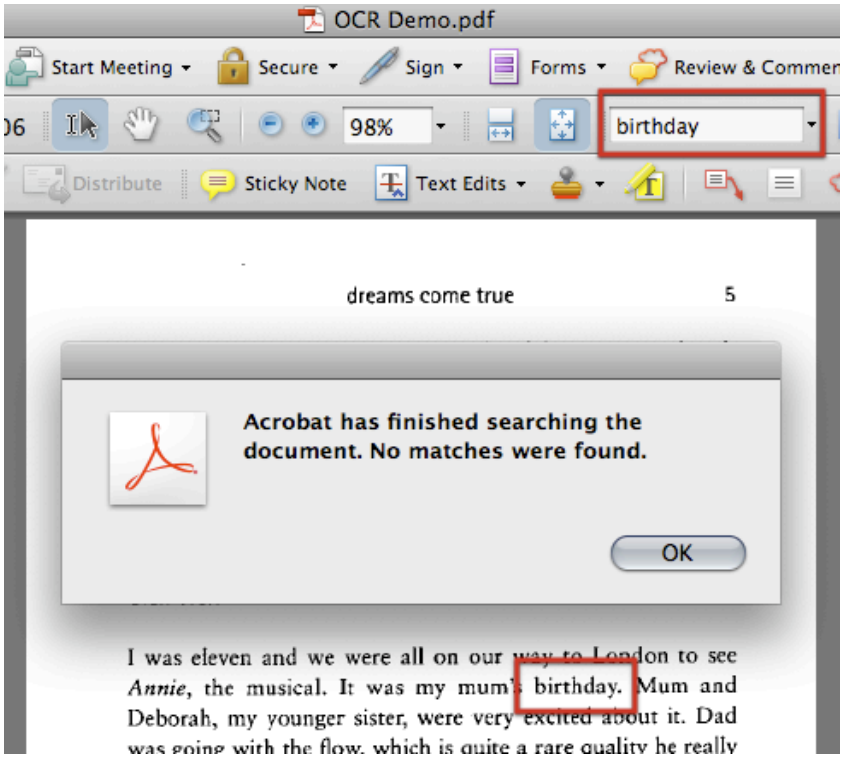
You should set your scanner to scan at a resolution of 300dpi, and check for other settings that may affect the quality of images. You should decide whether to scan in colour or greyscale (black and white). For documents containing a lot of images and/or complicated formatting (columns etc) you may find that colour scanning is more effective as it aids document decomposition.

TWAIN Compatibility

You should check that your scanner is TWAIN compatible as most OCR software requires TWAIN compatibility. Information about TWAIN compatibility should be available on the scanner manufacturer website, and OCR software manufacturers may list TWAIN-compatible scanner models on their website.

After scanning your printed materials you will have a number of PDF image files. These files are - like photographs - copied images of the printed material and cannot be searched for text. If you open one of these PDF files and try to use the Find function, it will not return any results.

For example,



In order to convert PDF image files to searchable text, you will need to perform OCR.

1.2 Choosing OCR Software

In order to carry out OCR on your PDF images you will need to choose and install OCR software. Many software packages have demo or trial versions available for download, and you might also want to check whether your scanner came with any OCR software included.

Depending on the language you wish to scan, and the operating system you use (Windows or Mac), there are a number of different OCR software packages available. Mac users currently have fewer OCR software options available and so may want to consider using a Windows PC to do the OCR process. Alternatively you could use Parallels or Boot Camp to install Windows on your (Intel) Mac. You are advised to search online for the most up-to-date OCR software, as OCR technology (especially for non-alphabet languages) is continually improving. You should check the system requirements for each software package for compatibility. For example, some OCR software can only be used with *localised* editions of Windows.

With good-quality PDF images, English OCR software is able to recognize text with a very high degree of accuracy. However, particularly if you are using non-alphabet languages, you are advised to test a few OCR software packages and compare their performance. If you carry out the OCR process on a sample document using a number of different software packages, you can compare accuracy by searching for a number of words to see which software has been most successful in recognizing them. You can also export the text to see which software preserves the original format best.

Example
Four OCR software packages were tested with Japanese data. The OCR output accuracy was compared, then the exported text formatting of the top-performing two software packages was compared.

OCR Software	Test 1: Accuracy Ranking	Test 2: Formatting Ranking
A	1=	2
B	3	
C	2	
D	1=	1

Software A and D performed equally well for character recognition accuracy, but software D was able to maintain the original text formatting more successfully. On the basis of the two phases of comparison, software D was selected for use.

1.3 Extracting Text

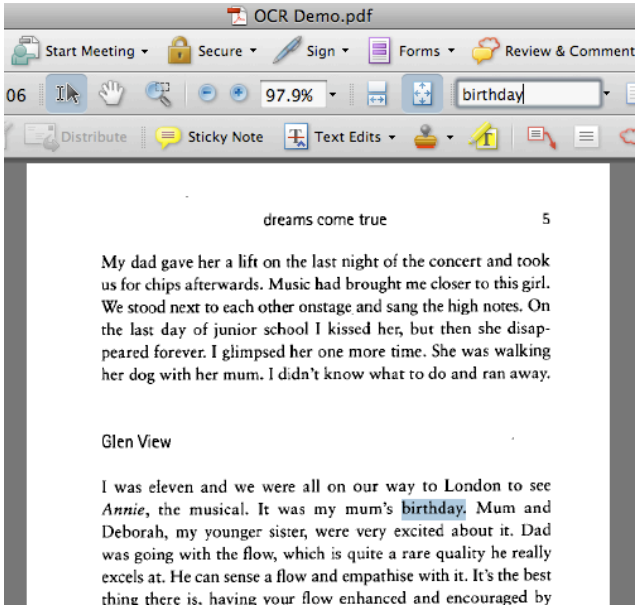
Once you have selected the OCR software you will use, you can carry out OCR on your PDF images.

Different OCR software packages have different settings and options. You should familiarize yourself with various settings to achieve the best results. If there is an option for "Autocorrect for skew" (which corrects minor rotations to pages after scanning) you should check this. Make sure you have set the language correctly before carrying out OCR.

When doing OCR on documents with complicated layouts (such as newspapers or scientific articles), consider using the Crop tool to select sections, and perform OCR section by section. This can help avoid formatting problems. As mentioned above, you may also find that scanning in colour can assist with successful document decomposition.

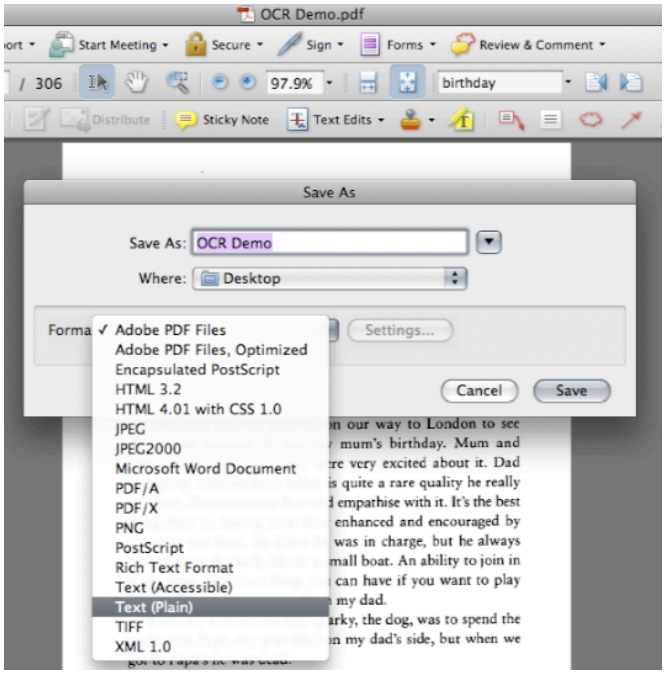
Once you have carried out OCR on your PDF images you will be able to search them for text using the Find function. Try searching for a word that you know appears in the text. When you save a copy of the PDF file after carrying out OCR, the PDF is now **searchable**.

e.g.

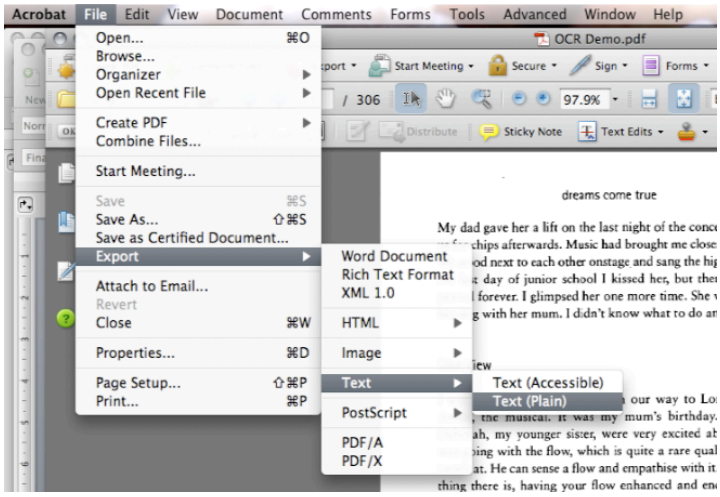


In order to carry out the next stage of the corpus-building process you need to use **searchable and editable** text. To do this, you should look for options in your OCR software that allow you to either:

(1) **Save As** a Word or text document



or, (2) **Export** the data as text.



After you have saved the text as a separate file (.doc or .txt or similar) you will probably have three files for each of your texts:

- a PDF image
- a searchable, OCRd PDF
- a searchable and editable text file

For the next stage of the process, you will need to use these text files, which are now your raw data.

2. Preparing Text and Header Files

This section will help you turn your raw data into text files that are suitable for use with the TEC Corpus Software.

2.1 Filenames

You will need to consider what system of filenames is most useful for your purposes. The conventions you adopt for naming files in your corpus will probably be related to the design of your corpus.

It is possible to identify types of texts (i.e. sub-corpora) using a system of filenames that indicates the text type. For example, in the TEC corpus fictional texts all share the letters **fn** in their filename, and biographical texts share **bb**.

Since the TEC Tools software was originally designed for use with a corpus comprising only texts translated into English, if you are building a comparable corpus that includes both non-translated and translated texts, or a parallel corpus of source and terget texts, you may wish to devise a system of filenames that helps you identify the source and target language of your translated texts, and/or differentiate between translated and non-translated texts.

Example

If a corpus includes English texts and their Japanese translations, and Japanese texts and their English translations there will be four types of texts. In such a case, texts can be given filenames such as:

English original	EE01	Japanese translation	EJ01
Japanese original	JJ02	English translation	JE02

These filenames allow the identification of sub-corpora - translated vs. non-translated texts (determined by whether the letters denoting language are the same or not), Japanese or English texts (the second letter is the language of the text) - and identify ST-TT pairs (the filename number identifies pairs of texts).

e.g.

Filename EJ07	Filename JJ04
Translated text	Non-translated text
Source Language = English	Language = Japanese
Target Language = Japanese	
Source Text Filename= EE07	Target Text Filename = JE04

Other systems of filenames can also be developed to encode various relevant attributes.

If you are building a corpus for personal use or a corpus that does not contain various textual sub-types, the filenames you use may not be important. However, it may be useful to consider filename conventions before building your corpus.

It is also possible to adapt the existing TEC header files to tailor them to your specific purposes. See section 2.3.

2.2 Text Files and Header Files

A corpus is typically made up of text files and header files.

- **Text files** contain the actual data to be analysed.

- **Header files** contain meta-data such as the title of the text, author, publisher and any other features of interest.

Text files and header files come in pairs, and the filenames for text files and their associated header files should be the same. Text files can be identified by a **.xml** extension and header files have a **.hed** extension.

e.g.

Text File	Header File
bb00000001.xml	bb0000001.hed

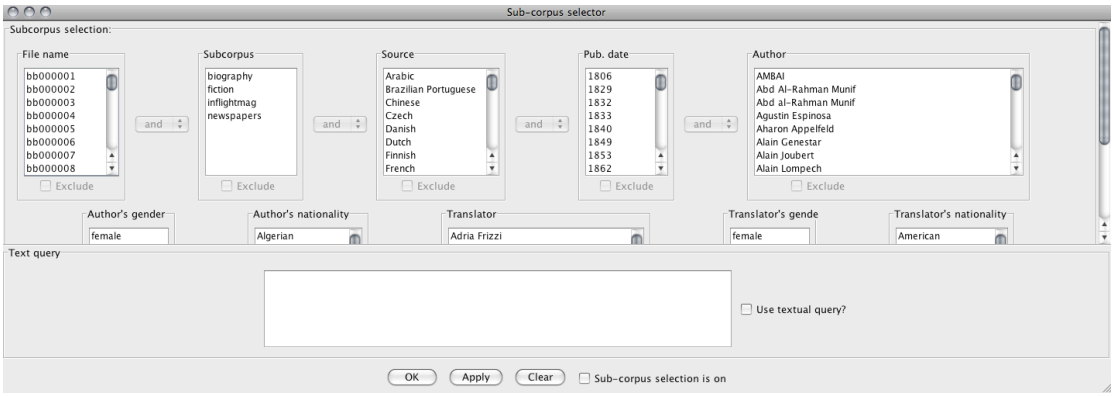
The TEC corpus and many other corpora use such a system of text files and associated header files because it allows the user to select sub-corpora - i.e. sets of texts that share certain features - to search in the corpus Browser.

Below is an example of a header file.

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<!DOCTYPE techeader SYSTEM "techeader.dtd">
<techeader>
  <title subcorpusid="biography" filename="bb000010">
    <subcorpus>biography</subcorpus>
    <collection>Girls of Alexandria</collection>
  </title>
  <section id="s1">
    <translator gender="female">
      <name>Frances Liardet</name>
      <nationality description="British"/>
      <employment>writer, translator</employment>
      <status>freelance; part-time</status>
    </translator>
    <translation extent="61739">
      <publisher>Quartet Books</publisher>
      <pubPlace>UK</pubPlace>
      <date year="1993">1993</date>
      <copyright>Frances Liardet</copyright>
    </translation>
    <translationProcess mode="wwst">
      <direction>into mother tongue</direction>
      <type>full</type>
    </translationProcess>
    <author gender="male">
      <name>Edwar al-Kharrat</name>
      <nationality description="Egyptian"/>
    </author>
    <sourceText>
      <language>Arabic</language>
      <status>original</status>
    </sourceText>
  </section>
</techeader>
```

You can see from this example that various pieces of information relevant to the text are listed. These include the name, nationality and gender of the author and translator, and the title, publisher and language of the text.

This information can be used to define sub-corpora when browsing the TEC corpus, as you can see from this screen-shot of the Sub-corpus Selector window of the Corpus Browser interface.



Some of the information in the header file can be input freely, for example, the name of the author or the title of the book. However, some information must be selected from a pre-determined list of options, for example, gender must be either 'male' or 'female'.

The rules that determine what can and cannot be included in a header file are defined in a file called a **DTD file**.

2.3 DTD Files

The rules that define what constitutes a "well-formed" text file or header file are listed in DTD files. When the TEC Tools software processes your text files and header files it needs to have access to an accompanying DTD file for the text files and another DTD file for the header files.

When you download the TEC Tools software, two DTD files are included in the folder: **tectext.dtd** and **techeader.dtd**.

After you have prepared all of your text files you will need to save them in a folder that also contains the text DTD file. Similarly, all of your header files must be saved in a separate folder that also contains the header DTD file.

If you want to change the information listed in the **techeader.dtd** file it is possible to do so if you are familiar with XML encoding. If you adapt the existing header DTD you can include any metadata that is relevant to your research. (For more information on DTD files, see Harold and Means 2002)

2.4 Setting up jEdit

In order to prepare both text and header files, you will need to use text editing software that can be used for XML encoding. jEdit is a freely available open source text editor that works on both Windows and Mac operating systems, among others. To start using jEdit, download and install the software (http://www.jedit.org/). The first time you use jEdit you will need to set it up. To do this, follow the following steps:

- Open jEdit

- Go to Utilities > Global Options > Encoding
- Select "Default Character Encoding UTF-8"
- Deselect "Auto-detect file encoding where possible"

- Go to Utilities > Global Options > Editing > Word Wrap
- Select "Soft".
- Click on Apply and OK.

- Go to Plugins > Plugin Manager > Install > XML
- Click Install
- Close

- Go to Plugins > Sidekick
- Check "Parse on keystroke"
- Check "Highlight markers in structure browsers"

By doing this you are preparing jEdit to work with XML encoding, and enabling jEdit to check for Errors in your files.

2.5 Preparing Text Files

Although there is considerable variation in the time it takes to prepare text files, based on the quality of the OCR text output and other variables, you might expect to spend between 1-2 hours to prepare the text from a 200-page book.

- Open your text file in Word.
- Do a spell check.
- Notice any repeated OCR errors (e.g. “Tm” for “I’m” etc) and correct these using Find and Replace.

As the text files will later be tagged with XML encoding - which enables the corpus software to process the text - you will also need to remove certain characters from your text, as they have specific meanings in XML.

Using Find and Replace,

- Find all instances of “<”and Replace them with nothing.
- Find all instances of “>”and Replace them with nothing.
- Find all instances of “&” and Replace them with “and”.

Then, **save the file as a text file with UTF-8 encoding**.

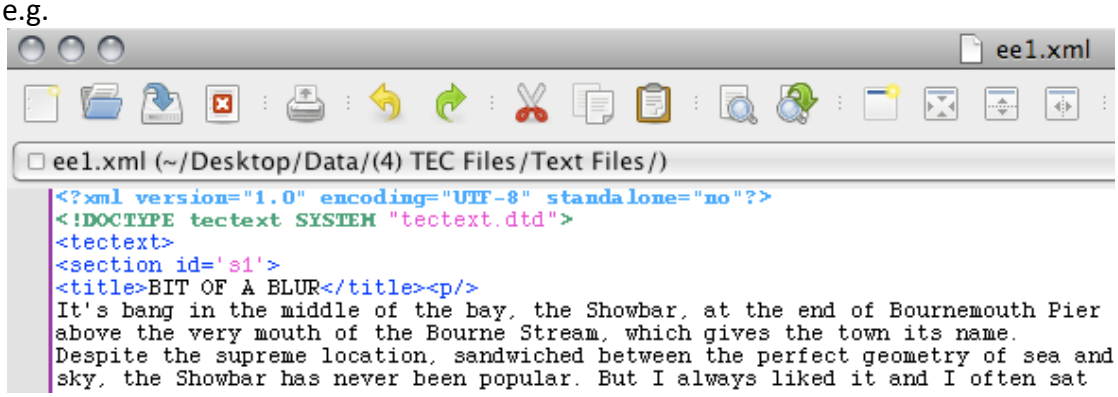
There are many ways to tag your text with information, and you should consider what information you will need in order to analyse your corpus. The most basic kind of XML tagging for text files is described below.

- Open the UTF-8 text file in jEdit.
- Remove any white spaces (i.e. delete blank lines)
- Either delete parts of the text that you do not wish to analyse (e.g. chapter headings, footnotes etc) or use <omit> tags to tell the Corpus Browser not to include this material in searches.
- Add the following XML tagging at the start of the body of text:

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<!DOCTYPE tectext SYSTEM "tectext.dtd">
<tectext>
<section id='s1'>
<title>INSERT TITLE</title><p/>
```

- Replace INSERT TITLE with the title of your text.

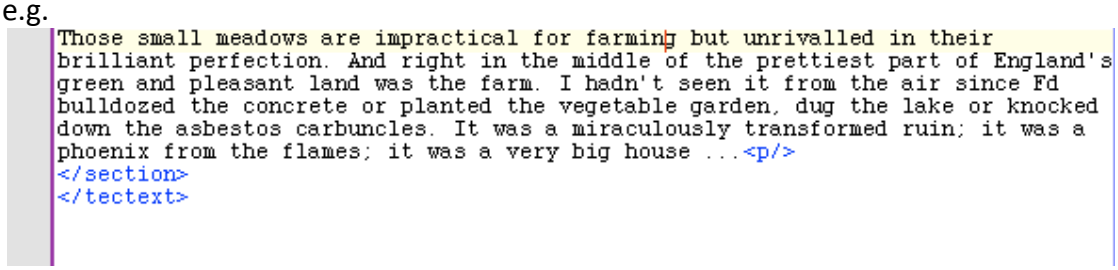
e.g.



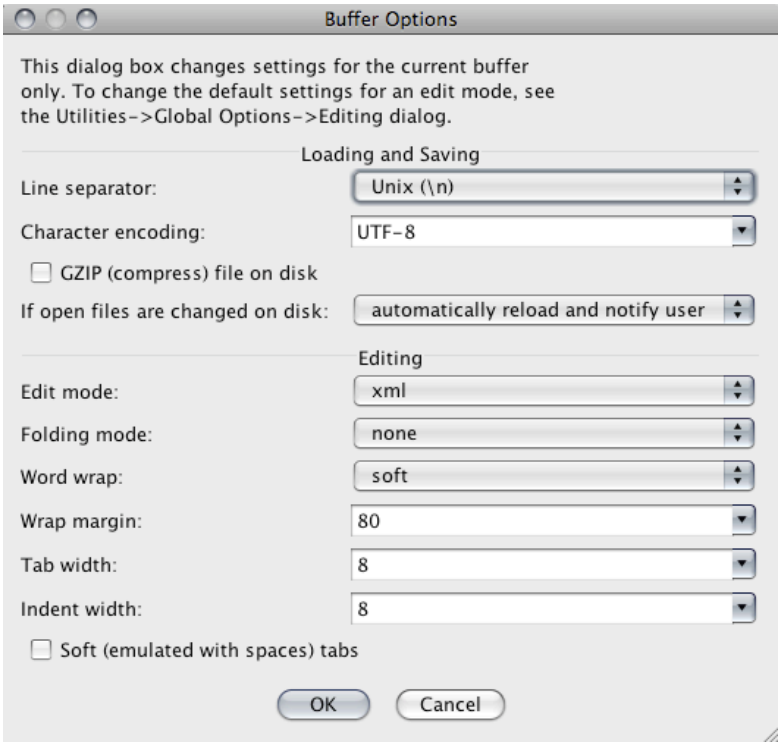
- Add the following XML tagging at the end of the text:

```
<p/>
</section>
</tectext>
```

e.g.



- Go to Utilities > Buffer Options. Use the following settings:



- Save and close jEdit.
- Put your text file in a folder with the text DTD file.
- Reopen the text file in jEdit.
- Check for errors by going to Plugins > Error List > Errorlist.

In order to ensure that you follow the same procedure for all texts, you may decide to create a **checklist** for preparing texts. Although there are many variations on the process for preparing texts for use with TEC tools, the very basic method outlined in this tutorial has been used as the basis for the checklist below.

NB.

- When saving files in jEdit you may find that duplicate files (with a ~ after the extension) are automatically saved. These are simply saved archive versions of your work.
- Although POS (Part-of-Speech) tagging is not currently available with TEC Tools, you may be able to use a separate POS tagging software that adds XML tags, prior to importing your text into jEdit.
- As jEdit is an open source software it may have bugs which affect its stability. Most issues can be resolved by closing down the software and reopening it, and it may be that you do not have any problems at all. However, you are advised to save your work regularly when using jEdit.

• Sample Checklist for Preparing Text Files

Filename e.g. ej01													
(1) Adobe Acrobat													
Export text													
(2) MS Word													
Spellcheck text													
(3) jEdit – Text Preparation													
XML	Find & replace < with nothing												
	Find & replace > with nothing												
	Find & replace & with “and”												
Check for Repeating OCR Errors	Find & replace ^ with nothing												
	Find & replace _1_ with “_I_”												
	Find & replace _!_ with “_I_”												
	Find & replace “TTi” with “Th”												
	Find & replace “TU_” with “I’ll_”												
	Find & replace “Td_” with “I’d_”												
	Find & replace “Tm_” with “I’m_”												
	Find & replace “T_” with “I_”												
	Find & replace “iH” with “ill”												
	etc …												
Working down the text	Delete or tag front matter up to start of text												
	Delete or tag headings, titles etc												
	Remove blank lines												
	Check for other errors												
	Repair broken lines												
	etc …												
	Save regularly (as bookname.txt)												
(4) jEdit – XML Tagging													
Paste title tagging from sample file													
Change title info etc													
Paste end tagging from sample file													
Set Utilities > Buffer Options > Line Separator > “Unix (\n)”													
Save As (filename.xml)													
Close jEdit													
Put in folder with tectext.dtd													
Open the file again													
Check for errors: Plugins > Error List													

2.6 Preparing Header Files

Header files are also prepared using jEdit. When you download the TEC Tools software folder, you will find some sample header files in the **eph** sub-folder. You can use one of the sample header files as a model for your own header files, simply by replacing the information in it.

Below is a sample header file.

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<!DOCTYPE techeader SYSTEM "techeader.dtd">
<techeader>
  <title subcorpusid="biography" filename="bb000010">
    <subcorpus>biography</subcorpus>
    <collection>Girls of Alexandria</collection>
  </title>
  <section id="s1">
    <translator gender="female">
      <name>Frances Liardet</name>
      <nationality description="British"/>
      <employment>writer, translator</employment>
      <status>freelance; part-time</status>
    </translator>
    <translation extent="61739">
      <publisher>Quartet Books</publisher>
      <pubPlace>UK</pubPlace>
      <date year="1993">1993</date>
      <copyright>Frances Liardet</copyright>
    </translation>
    <translationProcess mode="wwst">
      <direction>into mother tongue</direction>
      <type>full</type>
    </translationProcess>
    <author gender="male">
      <name>Edwar al-Kharrat</name>
      <nationality description="Egyptian"/>
    </author>
    <sourceText>
      <language>Arabic</language>
      <status>original</status>
    </sourceText>
  </section>
</techeader>
```

This header file has been created using the rules defined in a DTD file called **techeader.dtd**. The header file contains information corresponding to a text file called **bb000010.txt** in the biography sub-corpus of TEC, and is named **bb000010.hed**. Some of the information stored in the header file can be freely input (e.g. name of translator, title of book) whereas other information must be chosen from the list specified in the DTD file (e.g. gender must be male or female). If you open **techeader.dtd** in jEdit you will see the possible options listed.

NB.

- You do not have to enter all of the information listed in the header file, but this will limit the criteria you can use to search your corpus and define sub-corpora in the Corpus Browser.
- "Collection" refers to the title of the book.
- You do not need to enter a value for "Translation extent"

3. Building a Corpus

This part of the tutorial introduces the process of downloading and using the TEC Tools corpus software (also known as the MODNLP corpus suite).

3.1 Introduction to TEC Tools

The TEC Tools corpus software is a set of corpus tools that was developed by Dr. Saturnino Luz for use with the Translational English Corpus (http://ronaldo.cs.tcd.ie/tec2/jnlp/) and designed to allow free access to linguistic material over the internet. However, the software can also be downloaded and used to create and search a corpus using any texts you like. Originally the tools supported only English (and other European languages using Latin alphabet) but are currently being developed to allow use with non-alphabet languages such as Japanese.

The software consists of three modules:

- An **indexer** (called modnlp-idx) which allows you to create an index.
- A **corpus browser** (called modnlp-teccli) which can be used to select sub-corpora (by accessing indexes) and browse concordances.
- A **corpus server** (called modnlp-tecser) which can be used to make data and concordances (although not necessarily full texts) available to other users over the internet.

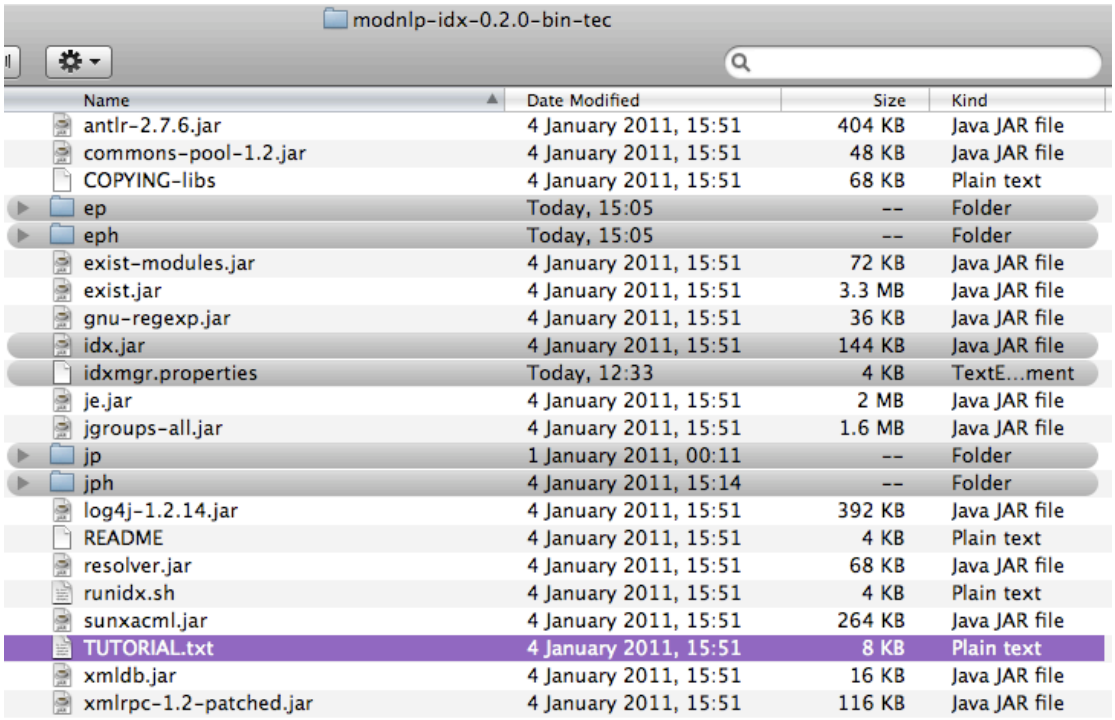
System Requirements

- The MODNLP software works with both Windows and Mac operating systems.
- You should ensure that your computer is using the most up-to-date version of Java available for its operating system.

3.2 Installing the Indexer

- Download the Indexer (IDX) which is available here: http://ronaldo.cs.tcd.ie/~luzs/tmp/modnlp-idx-0.2.0-bin-jp.tgz

- Unzip the folder to access the following files:



The **TUTORIAL.txt** file, highlighted purple, is the document on which the current instructions are based.

Other files and folders that you need to use are highlighted in grey.

- The **idx.jar** file is the **indexer**.
- The **idxmgr.properties** is a file that contains **settings** that are used by the indexer.
- The folders **ep** and **eph** are for storing English **data** files.
- The folders **jp** and **jph** are for storing Japanese **data** files.

3.3 Adding Data

➢ In the **modnlp-idx-0.2.0-bin-tec** folder, create a folder called **data**.

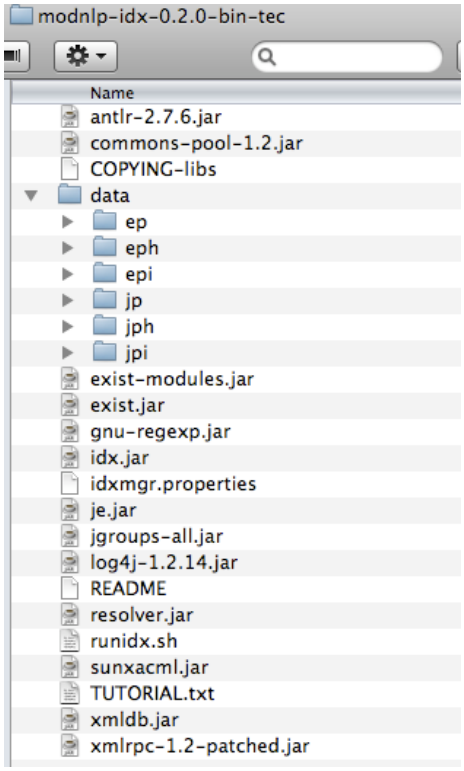
(a) Adding English-language data

➢ Move the folders **ep** and **eph** into this **data** folder.
➢ When you have prepared your **text files** (.xml files), put them into the **ep** folder, which already contains a dtd file called **ep.dtd** (and a couple of sample text files).
➢ When you have prepared your **header files** (.hed files), put them into the **eph** folder which already contains a dtd file called **ep.dtd** (and a couple of sample header files).
➢ Create another folder called **epi** and put this into the **data** folder too. This will be used later to store index files.

(b) Adding Japanese-language data

- Move the folders **jp** and **jph** into the **data** folder.
- When you have prepared your **text files** (.xml files), put them in the **jp** folder which already contains a dtd file called **tectext.dtd** (and a couple of sample text files).
- When you have prepared your **header files** (.hed files), put them in the **jph** folder which already contains a dtd file called **techeader.dtd** (and a couple of sample header files).
- Create another folder called **jpi** and put this into the **data** folder. This will be used later to store index files.

If you have added both English and Japanese data, you should have the following sub-folders in your data folder:

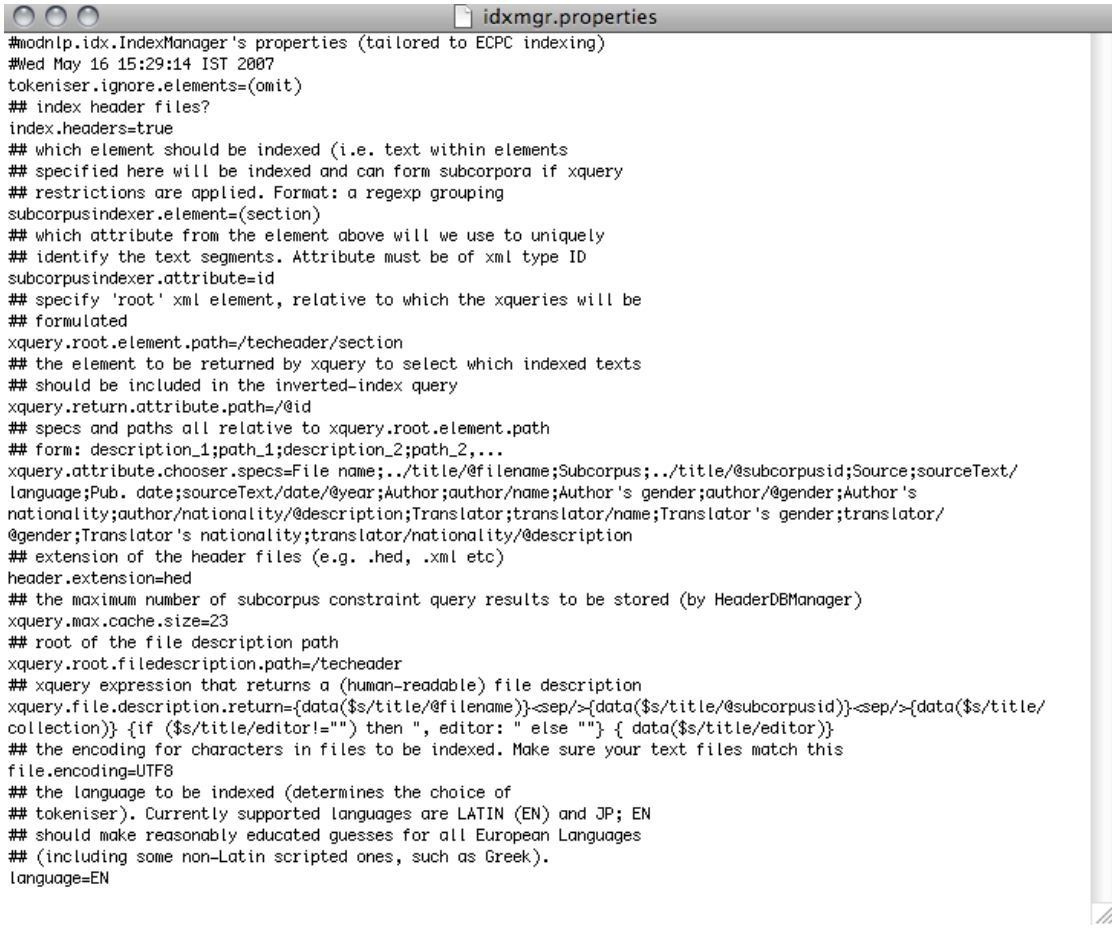


NB. The data folders **ep** and **eph** (**jp** and **jph**) already contain **.dtd** files and some sample text and header files. You may prefer not to replace these with your own data until you have completed the installation and testing process. If you have successfully tested the Corpus Browser using the sample data provided, you will be familiar with the procedure. If you then encounter problems when using your own data, troubleshooting may be more straightforward. When you have added your own data files to the data folders, remember to remove the **sample** files that were there from the start, but leave the **.dtd** files.

3.4 Setting up in the Indexer

idxmgr.properties tells the index manager how to tokenise text files, which sections of the text to index, and which parts of the corpus should be included in a sub-corpus etc.

The idxmgr.properties file can be opened using any text editor.



In this version of the software it is possible to set the indexer to work with English (EN) or Japanese (JP) text files that are encoded in UTF-8.

When the setting is EN the indexer should also be able to handle European languages. (See screenshot bottom 5 lines.)

If you want to build a corpus that contains both English texts and Japanese texts, you will need to change the **language setting** between JP and EN when indexing each set of data files. You should complete the indexing process for your English-language data, and then repeat for your Japanese-language data. When you carry out the indexing process, you should check the language setting is correct for your current data.

To change the language setting:

- OPEN idxmgr.properties (using TextEdit or any other text editor)
- Go to the bottom line that says 'language=JP' or ‘language=EN’.
- Change the setting to 'language=EN' or 'language=JP' as you require.
- SAVE and CLOSE

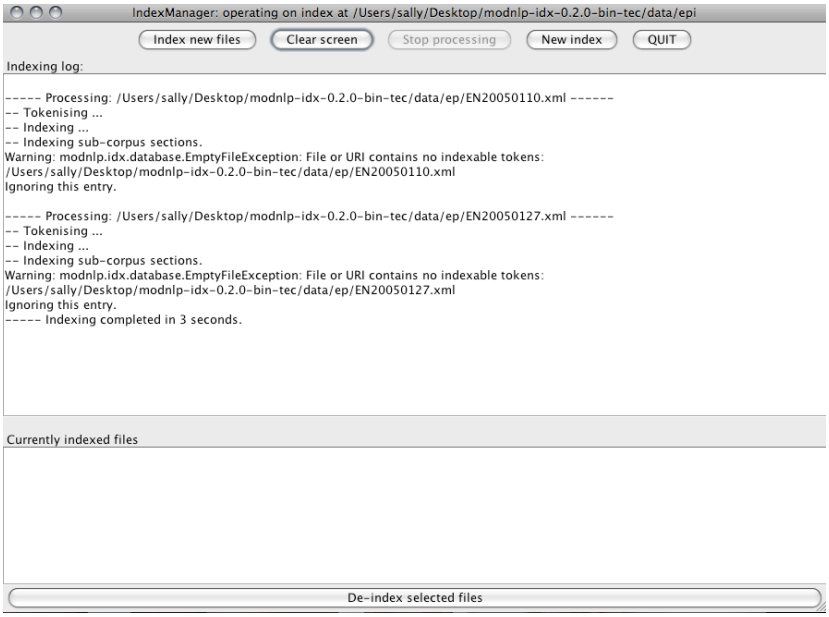
3.5 Indexing Files

- Open **idx.jar**. (NB. It may appear as just **idx**).
- A window will appear asking you to "Choose a location for the index". This location is where the Indexer will store the index it creates. The index will later be used by the Corpus Browser.
- In this window, open the **data** folder and select (i.e. highlight but do not open) either sub-folder **epi** or **jpi** (depending what set of data you are currently indexing).
- Click "Choose a location for the index".
- Next you will be asked to choose the folder where header files are stored. Select **eph** or **jph**.
- After about 30 seconds pause you will be asked to give a URL for public access to the header files. This is relevant if you want to make your corpus available to other users. Click "OK" to accept the default URL.
- The main indexer window should now appear. Click "Index New Files". Open the **ep** or **jp** folder and select the XML files you want to index. Click "Index New Files".
- When the indexing process is finished, click QUIT > Yes.

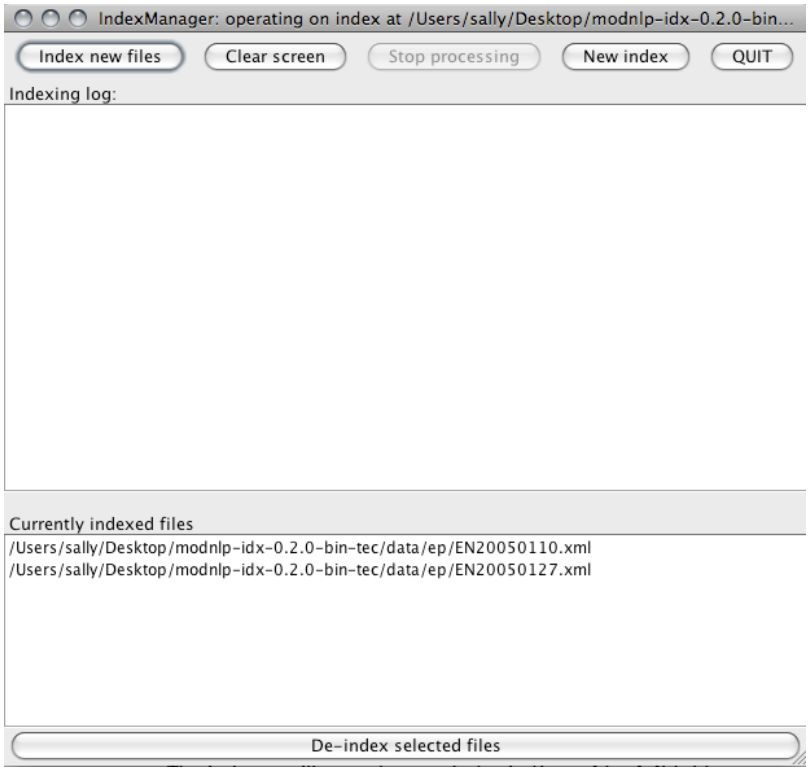
The indexer will have stored an index in the **epi** or **jpi** folder.

Troubleshooting:
Check that you have the correct language setting in the **idxmgr.properties**.
Check that you *selected* but did not *open* folders.

If you get an error message such as this:



Close the Indexer and reopen it. You may find that the issue has been resolved.

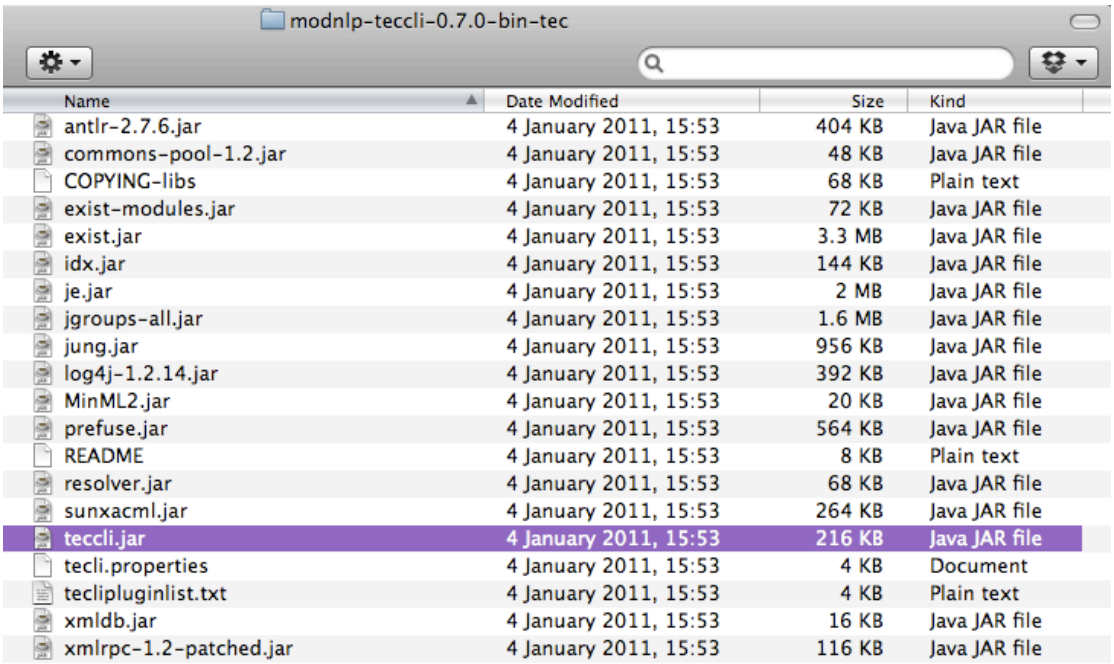


8. Using the Corpus Browser

To search your corpus you will need to use the Corpus Browser (TECCLI). The Corpus Browser works with the index created by the Indexer (IDX) as described above to select and search your texts.

- Download the Corpus Browser (TECCLI) which is available here: http://ronaldo.cs.tcd.ie/~luzs/tmp/modnlp-teccli-0.7.0-bin-tec.tgz

- Unzip the folder to access the following files:



- Click on **teccli.jar** (highlighted above). (NB. It might appear as just **teccli**.)
- A window will appear asking you to select a corpus. The default corpus is TEC, which is available on-line. To view your own corpus, select "Choose new local corpus".
- A window will open that asks you to "Choose location for the index". Select the folder that the index is saved in, i.e. the **epi** (or **jpi**) folder.
- A window may open that asks you to "Choose a headers directory". Select the folder that the header files are saved in, i.e. the **eph** (or **jph**) folder.
- The concordance browser window will appear.
- Type "*this*" to check the concordancer is working.

Troubleshooting

If you have followed these instructions but the concordance browser does not launch, try the following:

- Check what versions of Java you have installed.

For Mac OS X users:

Open Utilities > Java Preferences > General tab.
Move the most recent version to the top of the list.

- Check which version of Java you are using.

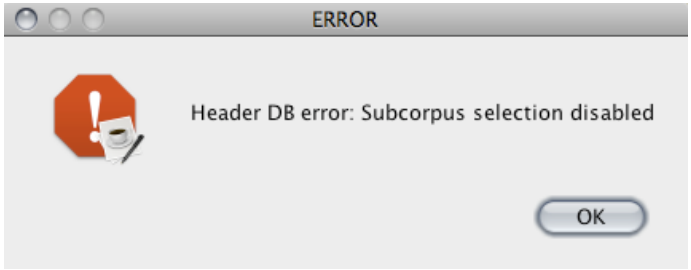
For Mac OS X users:

Open Terminal (this can be found in Applications > Utilities > Terminal)
Type "`java -version`", then Enter

- Confirm that this version is compatible with the current version of the Corpus Browser.

- Check your Java security settings to confirm that you are not blocking anything (Utilities > Java Preferences > Security tab).

If the concordance browser launches but gives this error message:



Click **OK** and proceed.

3.6 Sharing your corpus

If you want to make your corpus available to other users, you will need to download the Corpus Server module (modnlp-tecser).

4. Searching a Corpus

4.1 The Corpus Browser Interface

Many of the features of the Corpus Browser are self-explanatory but brief descriptions of some of the menu options are provided below.

File Menu

- **New local corpus**
  If you have saved a corpus on your computer, you can select it by selecting the folder where the index is stored, e.g. the **epi** folder.

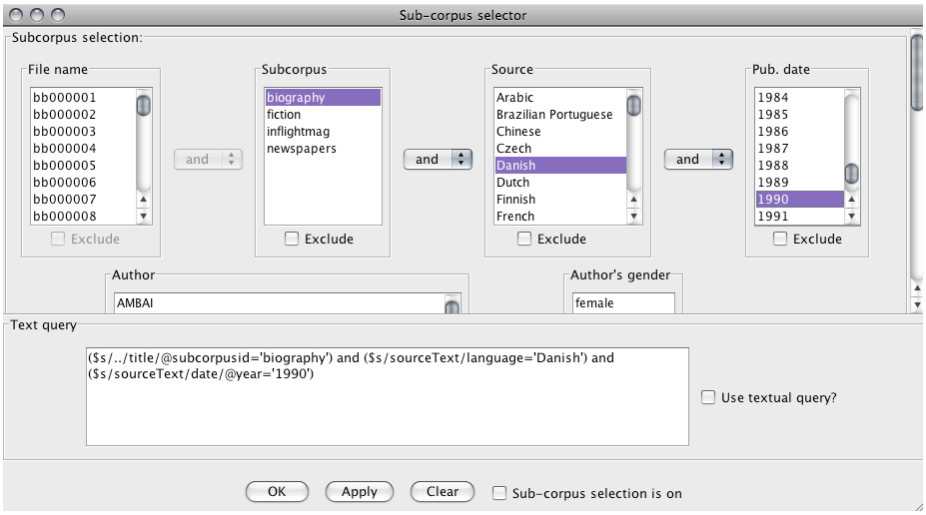
- **New Internet corpus**
  If you want to access a corpus remotely you should input the IP address of the server on which the corpus is stored. The default IP address provides access to the TEC internet corpus.

- **Save Concordances**
  You can save the results of your corpus searches. To do this, click "Save Concordances" > specify a filename and location > "Save to Disc". You can open the saved file with a text editor or Word. To set out the concordances clearly in Word, go to Page Setup > Orientation > Landscape. Then set the font to Courier size 8.

Options Menu

- **Case sensitive**
  The default setting for corpus searches is that they are not case-sensitive. If you want to carry out case-sensitive searches, check this option.

- **Select Sub-corpus**
  This option opens the "Select Sub-corpus" window. In this window you can select particular text files or groups of text files to search. There are a number of selection options including date of publication, gender of translator, source text language etc. The options available are based on the information stored in the header file for each text in the corpus.



You can specify an AND / OR relationship between the attributes you select. You can also check the "Exclude" box to indicate a NOT relationship.

The simplest way to select a sub-corpus is to highlight attributes listed in the dialogue box. However, it is also possible to write specific queries using the "Text Query" box at the bottom of the window. By writing queries in this text box, it is possible to specify various combinations of metatextual information stored in header files - including information that is not listed in the sub-corpus selection boxes (e.g. Publisher) - with various logical operators. If you wish to develop your own header DTD file, including information not listed in the DTD file supplied with TEC Tools, you may wish to make use of the Text Query function in order to operationalise the metadata you have chosen to include.

When using the Text Query box:

- Use XQuery syntax
- After writing your query, check the "Use textual query?" box

If you are not familiar with using XQuery, you may wish to consult the relevant chapter in "XML in a Nutshell" (see reference in section 5). Alternatively, you can try defining various sub-corpora by highlighting attributes in the lists, and observe the text query your selection generates. You can then adopt the same conventions to write other queries.

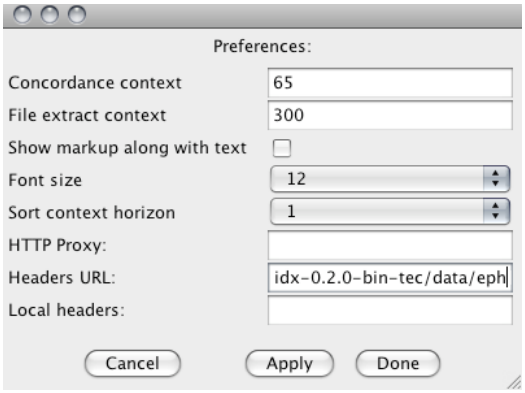
e.g. Using the menus I can create a query for "*fictional texts translated from Chinese, where either the author or translator is female, but is not Zhisui Li*" by highlighting items and using the AND/OR and EXCLUDE options.
The text query syntax this automatically generates is:

```
($s/../title/@subcorpusid='fiction') and
($s/sourceText/language='Chinese') and (not ($s/author/name='Zhisui
Li')) and ($s/author/@gender='female') or
($s/translator/@gender='female')
```

- **Activate sub-corpus selection**
  This option activates and deactivates the selection criteria set in the "Select Sub-corpus" menu. When this option is deselected, any searches will be carried out on the whole corpus.

Preferences

- This option opens the "Preferences" window in which you can change various settings.



Some settings you may wish to experiment with include:

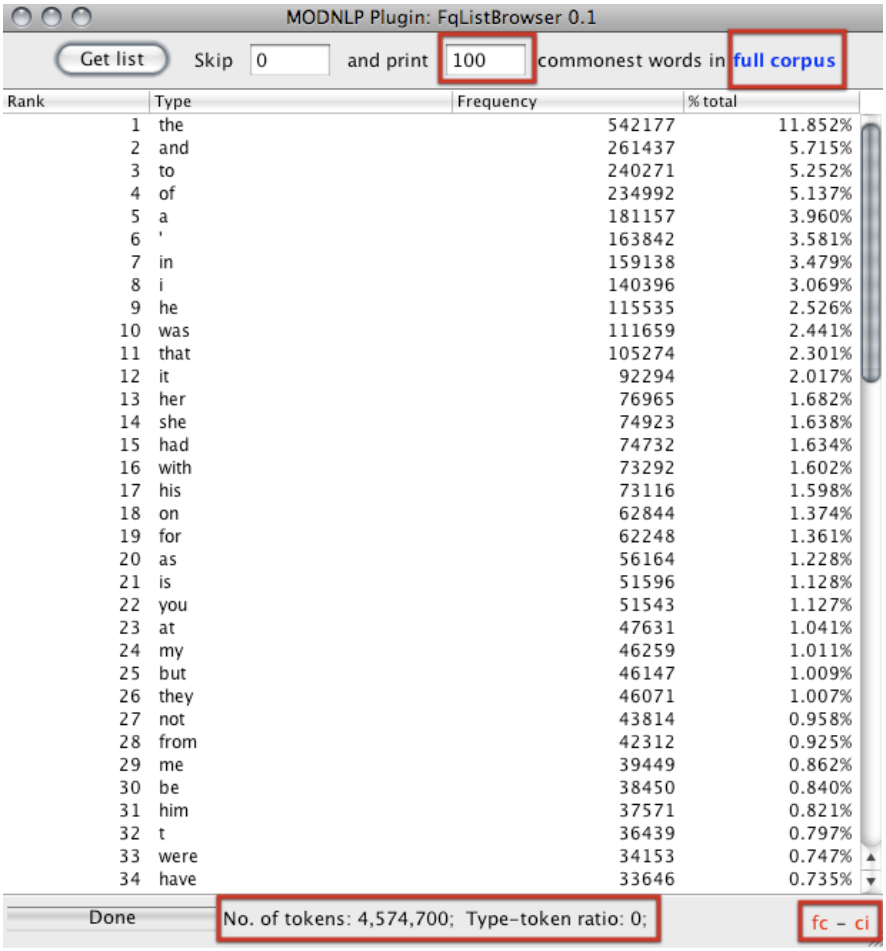
Concordance context > changes the number of words that are shown in concordance lines.

File extract context > changes the number of words that are shown when using the "Extract" function to check individual concordances in their broader context.

Show markup along with text > shows you any xml markup that is present in the text file. For example, </p> indicating a paragraph break.

Plugins Menu

- **Word frequency list**
  This tool generates a frequency list for the corpus. You can change the number of words listed (e.g. 100 most frequent words, or 500 most frequent words) and click "Get List". You can save lists which can later be opened using a text editor or Word.

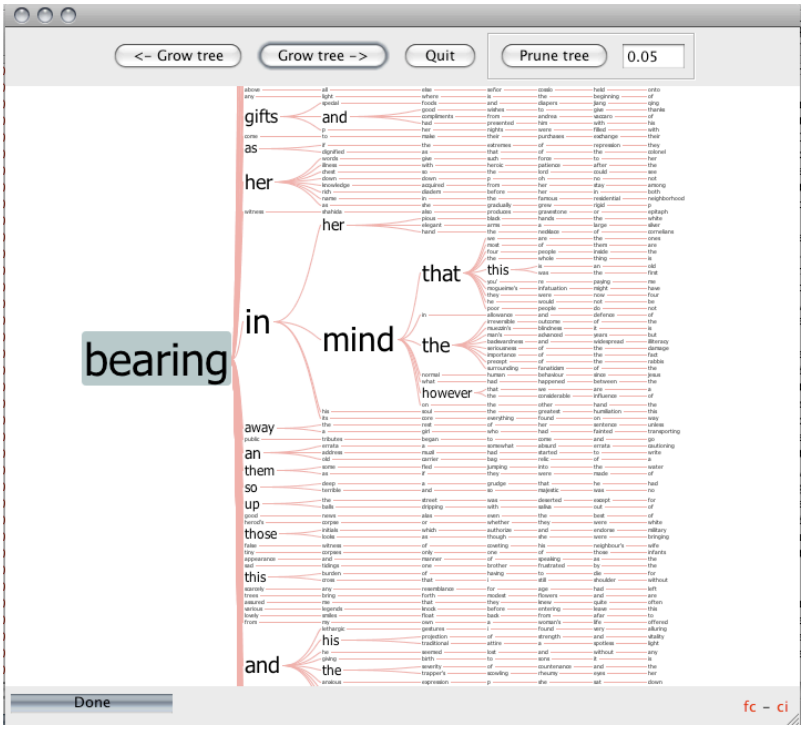


Rank	Type	Frequency	% total
1	the	542177	11.852%
2	and	261437	5.715%
3	to	240271	5.252%
4	of	234992	5.137%
5	a	181157	3.960%
6	'	163842	3.581%
7	in	159138	3.479%
8	i	140396	3.069%
9	he	115535	2.526%
10	was	111659	2.441%
11	that	105274	2.301%
12	it	92294	2.017%
13	her	76965	1.682%
14	she	74923	1.638%
15	had	74732	1.634%
16	with	73292	1.602%
17	his	73116	1.598%
18	on	62844	1.374%
19	for	62248	1.361%
20	as	56164	1.228%
21	is	51596	1.128%
22	you	51543	1.127%
23	at	47631	1.041%
24	my	46259	1.011%
25	but	46147	1.009%
26	they	46071	1.007%
27	not	43814	0.958%
28	from	42312	0.925%
29	me	39449	0.862%
30	be	38450	0.840%
31	him	37571	0.821%
32	t	36439	0.797%
33	were	34153	0.747%
34	have	33646	0.735%

To reveal the meaning of the abbreviations - e.g. **fc** (full corpus) and **ci** (case insensitive - appearing in the bottom right corner, hover over them with the cursor.

- **Corpus description browser**
  This lists the filenames, title and other information (meta-data) for the texts in the corpus. The number of tokens and type/token ratio for each text are also listed.

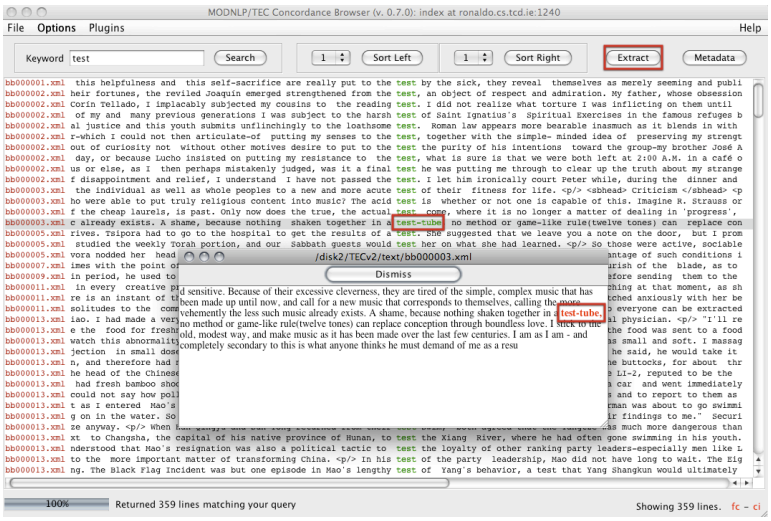
- **Concordance tree viewer**
  This is a corpus visualization tool that enables you to “grow” a tree for various keywords. The tree view can help you to identify frequently occurring collocations, since the size of the text reflects frequency of occurrence in the corpus.



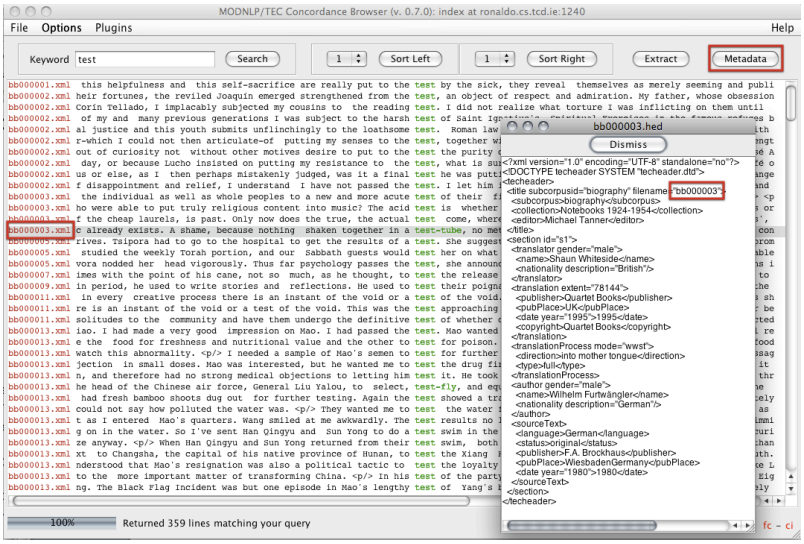
- **Concordance mosaic viewer**
  This is another corpus visualisation tool.

Other Tools

- **Extract**
  Allows you to see more context for a particular concordance. Select the concordance line by clicking on it and then “Extract”.



- **Metadata**
  Allows you to see the information stored in the header file for a particular concordance. Select the concordance line by clicking on it and then “Metadata”.


MODNLP/TEC Concordance Browser (v. 0.7.0): index at ronaldo.cs.tcd.ie:1240
File
Options
Plugins
Help
Keyword
test
Search
Sort Left
Sort Right
Extract
Metadata
bb000003.hed
Dismiss
Returned 359 lines matching your query
100%

4.2 Searching a Corpus

The main types of search query types are listed here. This information is taken from the Help file which is accessible through the Corpus Browser:

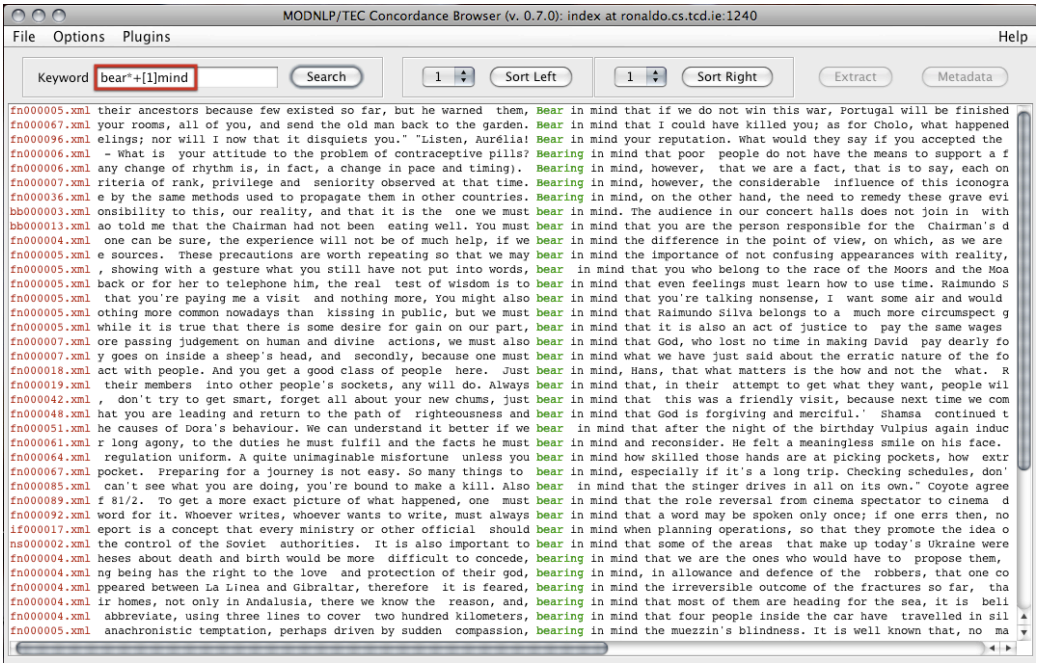
- **Single Keyword**
  This kind of search will return all instances of a word in a corpus (or selected sub-corpus). The default search setting is not case-sensitive. If you want to specify that searches should be case-sensitive, check Options > Case Sensitive.
  e.g. seem > seem, Seem, SEEM

- **Wildcards**
  Wildcards allow you to search for all words beginning with a certain prefix, or ending with a certain suffix. ***** denotes a wildcard.

  e.g. seem***** > seem, seems, seemed, seemingly
  *****hat > that, what

- **Sequences**
  You can search for sequences of words using a **+** sign. You can also specify the number of intervening words using **[#]**, and include wildcards ***** in your search.

  e.g. seen**+**before > never seen before
  know**+**before***** > know beforehand
  seen**+[1]**before > seen it before, seen her before
  seen**+[1]**before***** > seen her beforehand



4.3 Sorting Concordances

When you have generated concordance lines, you can arrange them in various ways to help you identify patterns. There are three ways to do this.

- **Sort Tool**
  Allows you to sort concordances to the **left** or **right**, and specify the number words between the search keywords.
  NB. In cases where a search returns only a small number of concordance lines you may find that the Sort buttons are not active. By reducing the size of the window so that a scroll bar appears on the right, the Sort buttons should become active.

- **Tree Viewer** (described above)

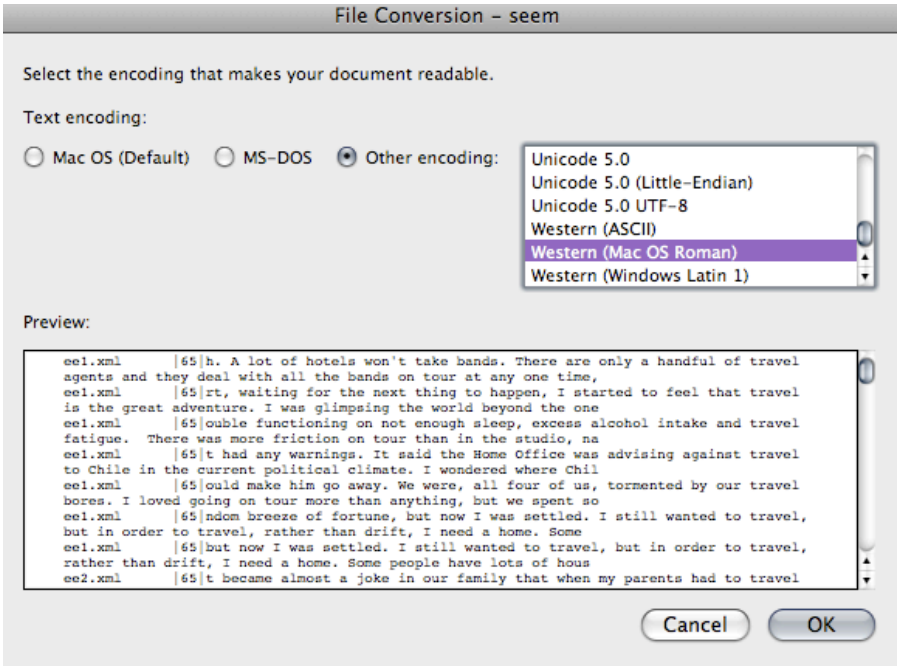
- **Mosaic Viewer** (described above)

4.4 Saving Search Results

Although it is not possible to print concordances directly from the Corpus Browser, you can save concordances, frequency lists, etc. You can then open the saved concordances for printing, annotation and editing later.

You can save concordance results (File menu > Save concordances) or frequency lists, and corpus descriptions (click Save button) which are available in the Plugin Menu. Then specify a filename and choose the location you wish to save to by double-clicking the location.

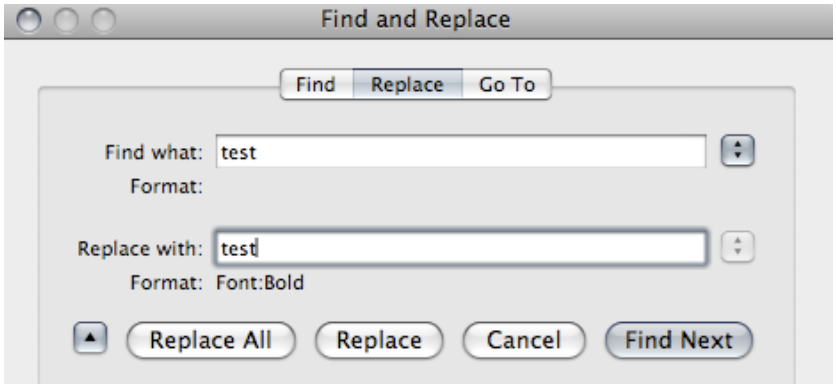
The saved file contains UTF-8 text which can be opened using a Text Editor or Word. When opening the file using Word you may be asked to specify the file encoding:



Any of the three text encoding options here can be used to successfully open the file.

To set out concordances clearly in Word:

- Go to Page Setup > Orientation > Landscape
- Set the font to Courier size 8 (or other fixed-pitch font)
- Find and Replace "`.xml |65|`" with " " (i.e. one space)
- Add line numbers for easy reference
- Add bold formatting to your concordance keyword using Find and Replace to add the bold font formatting.



5. Other Resources

Other resources you may wish to consult include:

http://www.monabaker.com/tsresources/TranslationalEnglishCorpus.htm
This website provides information about TEC and its contents, a select bibliography, and a PDF presentation by Dr. Saturnino Luz, introducing the TEC Tools corpus software.

http://ronaldo.cs.tcd.ie/tec2/jnlp/
This website provides access to the TEC resource using the Corpus Browser. You can use this to familiarize yourself with the Browser interface and try out searches as described in section 4.4.

http://modnlp.berlios.de/doc/
This website has documentation of the TEC Tools suite of software for developers.

There are a number of publications that you may find useful.

Luz, S. (forthcoming) 'Web-based Corpus Software' in A. Kruger, K. Wallmach and J. Munday (eds) *Corpus-Based Translation Studies: Research and Applications*, London: Continuum.

Harold, Elliotte Rusty and W. Scott Means (2002) *XML in a Nutshell*, Cambridge MA: O'Reilly Media.