Indexing with Xaira

Xaira – the XML version of Sara – is very much work in progress. This document is also only a very preliminary draft introduction.

Copy the file `xaira-texts.zip` from drive T: to your working directory `D:documenti`, and unpack it into a directory called `Xaira`. It contains a folder called Varney and another called Forli, which contains the texts you prepared yesterday. In this exercise we will work on the Varney files first, and then on your own files.

1 The bare minimum

We begin with the simplest case: a text file containing no markup other than a `<text>` tag at the start of the file, and a `</text>` tag at its end.Let's see how far we can get with this minimal approach.

The current version of Xaira is in the Linguistic Tools folder on your desktop. There are two icons: one to launch the Index Toolkit and one to launch the query client.

1.1 Indexing a plain text file

1. Open the Xaira Indexer Toolkit by double clicking on its S icon (it says "SARA" but that's just to confuse you).
2. Select New from the File Menu.
3. Select Parameter File from the Tools menu
4. In the Parameters dialog, type `varney` or some other name you like into the Name box. This is the name of our corpus. Now use the Browse button to select the folder `Xaira/varney` as the Root for this corpus.
5. Press the Defaults button. Xaira will obligingly fill in default names for all the other parameters. For the moment, we will accept these: Press OK
6. Now choose File List from the Tools menu. Xaira opens the filelist menu so that we can specify which of the files in our nominated text folder (`Xaira/varney/texts`) are to be indexed. Press the Generate button.
7. By default, both the files in the Texts folder appear. Select either the one you want (`varneyplain.xml`), or the one you don't want (`varney.xml`) and then press either the Select or the Delete button, respectively. Either way, the lower window should contain only the name of the file you want to index before you press OK
8. If we had a corpus header for this corpus, we could now proceed to validate it. Since we do not, we need to create one. Choose Make Header from the Tools Menu
9. Using Windows Explorer, look in the folder `Xaira/varney` again: you will see that it now contains a file called `corpus_header.xml`. Open this file and you will see that Xaira has inserted all the information it can about your new corpus: its name, and the tags it has found within it
10. Just for completeness, now select Parse all from the Tools menu. You should see the message `0 errors`. Later, we will see what happens if you try to build a corpus containing invalid XML files.
11. Now select Tag Usage from the Tools menu. A list of the tags found in your corpus is displayed, together with a count for how often they appear. You can edit this list to include a brief gloss for each tag, which will later be used by other parts of the system to remind you what its function is. Try this out, by selecting a tag, and pressing the Edit button.
12. If our corpus had more markup, we could make use of it, for example to provide more sensible references for individual hits, or to provide detailed bibliographic descriptions for each text. Since it does not, however, we will just save the current files, by selecting Save from the File menu.
13. We are now ready to index our corpus. Click on `Indexer` on the Tools menu, and choose `Run Indexer` from the submenu which opens. A new window appears, briefly, in which you will see the name of the files being indexed and various other diagnostic messages. When the process is complete, this command window will close itself: unless your files are very large, they should only take a few minutes to index.

14. We can now check to see what is in the index. Choose Test Index from the Test menu. Dismiss the dialogue warning you that you have not created a bibliography for your corpus by clicking on the OK button. The "View word" dialog appears.

    9

    - Type a word which you know appears in your corpus into the box at the top of the screen and press the `View` button. The next box down shows you the frequency with which this word appears (ignore the forms column for now: this text does not have enough markup for different word forms to be detected)
    - Select the word you typed from the frequency list by clicking on it: by default, versions of the word with different capitalization are distinguished by the indexer. The next box down is filled with a scrollable list showing the frequency of each different spelling of the word.
    - Select one of the spellings of your search word to see a list of all the places in which your corpus where this spelling occurs.
    - Click on the start of one of the lines in the occurrence list to see the context for this occurrence in the bottom window. Since we have no tagging in our document, the indexer has used each input line to define the context by default.
    - You can step through the contents of the index one at a time by clicking on the arrows towards the bottom of the occurrences pane. The other panes change to show occurrences of the items highlighted: select an item from the word forms pane to see all occurrences of it, as before.

1.2 Configuring the Xaira client

The Xaira client can access any corpus indexed by Xaira, once you have told it how to find the right corpus parameter file. To do so, proceed as follows:

1. Double-click on the Xaira icon in the Linguistic Tools folder: the Xaira splash screen appears.
2. Click the Menu button: the Server List dialog opens
3. Click the Add button: the Server properties dialog opens
4. Fill in the Server Properties dialogue as follows:]
    - Type the name of your corpus into the Name box (`minimal` if you are continuing the previous exercise)
    - Check the Local box, so that it has a tick in it
    - Click the Browse button and navigate to the folder containing your `corpus_parameters.xml` file (`xaira/varney`)
5. Select the file and press OK: the server list dialog reappears, with a new entry in it for the corpus you just defined.
6. Select the name of the corpus you want to open, and press the OK button to open it; alternatively, press the Set default button to make this the default corpus, so that the next time you start Xaira it will proceed to open this corpus without your having to go through the corpus selection procedure.
7. Now open the Word Query window and press Return. After a pause, all the different word forms in your file will be displayed. You can click on the column headings to sort by frequency, and scroll the list.
8. Click on a word that interests you, and then press the Query button to see how it is used in your corpus.
9. Xaira offers you many ways of changing the context displayed for each hit. You can expand the context, swtch to XML view, toggle between line mode and page mode, etc. It also offers you a wide range of different kinds of search facilities. These are all fully described later in this tutorial.
10. Remember to close Xaira down by selecting Close from the Files menu when you have finished your preliminary explorations of it.

1.3 A little more markup...

Adding a little markup to Varney would improve the situation. We could do this by editing the text or by pre-processing it automatically. To save time, we have prepared a better version of the corpus, by using an XSLT stylesheet to make the following changes:

- Each chapter is wrapped in a `<div>` element, which carries an `n` attribute giving the chapter number.
- Each paragraph (as indicated by two consecutive linefeeds in the original) is wrapped in a `<p>` element.

- The metadata (following a row of equals signs in the original) has ben removed from the text
- Each italicized phrase (bracketted by underscores in the original) is wrapped in a <hi> element.
- The first three paragraphs of each division are tagged as <head> elements.

If you open the file with an XML-aware browser such as Internet Explorer, you can visualize the structure of this XML file.

We will now re-index the file, taking advantage of the markup we have just introduced.

1. Double click on the Xaira Index Tools icon again.
2. Select Open from the File Menu and navigate to the `corpus_parameters.file` in the `xaira/varney` folder.
3. As before, select File list from the Tools menu. This time we want to include only the file `varney.xml`, so delete the unwanted file `varneyplain.xml` after pressing Generate. Press OK.
4. Since this is an XML file, we can now validate it. Select Parse all from the Tools menu. A message saying 0 errors appears, confirming that this file is valid XML.
5. As before, we need to generate a header: press Make Header, and then press the Save button, or select Save from the file menu. Xaira will write three XML files (the earlier versions will be overwritten, but that doesn't matter).
6. Open the file corpus_header.xml with your editor, or in IE5. You will see the default information that Xaira has written to the header, including the counts for the various elements found, and a default name for the corpus. We can use the Xaira tools to add more information to this corpus header for later use by Xaira.
7. Select Tag Usage from the Tools menu. A list of elements appears, with their counts, derived from the <tagUsage> element in the header.
8. Select the div element by clicking on its name in the list, then click the Edit button. A dialog appears in which you can enter a gloss: type in `a chapter` and press OK.
9. Now choose the References command from the Tools menu. This is used to tell Xaira how it should reference or label hits when you search the corpus. A reference label has two parts: one specifing the text and the other one or more units within that text, such as a paragraph or sentence. In the BNC, for example, a reference such as CCC 123 means sentence number 123 in corpus text CCC. The unit is `sentence`, the text is `corpus text`.
10. As you see, the default 'text' is a file, and the default 'unit' is a line-number. As all of our text is in a single file, the current 'text' reference is not much use to us. To change it, first select the `0` to the left of the `Text` line in the upper part of the References dialogue and press Edit.
11. For Varney, we will use the chapter numbers to identify each 'text'. These numbers were supplied on the N attribute of the element: to use them, select first `div` from the available list of element names, and then select `n` from the list available attributes. Finally, make sure that the radio button `Use this tag/attribute` is selected.
12. If you close this dialogue now, references will consist of the chapter number and the line number within the whole file. Let us add another unit: the paragraph number within the chapter. In our tagging the paragraphs do not have any explicit numbers, but Xaira can count them for us.
13. Click on the Unit line in the upper window, and press the Add button. Then select `p` from the scrollable list of element names, and click on the radio button labelled `Use this tag and compute label`. Finally, press OK, and then press the Save button again to save your changes.

1.3.1 Creating bibliographic information

As well as displaying a short reference for each hit, the Xaira client can display more detailed information about individual texts. This information is usually extracted from the headers of individual texts, but it can be derived from any part of the corpus using an XSLT stylesheet. We will use the <head> elements we added to the start of each in the Varney corpus for this purpose.

- Choose References from the Command menu
- Xaira displays the XSLT style sheet which is used by default to extract bibliographic references for each text. You don't need to know much about XSLT to understand this: just change the template to read `select="//head[2:4]"/>`: this means use its <head> elements as a bibliographic reference for every text.

- Now select Make bibliography from the Command menu and press OK: Xaira Tools will now create a file called `bib.xml` which the Xaira client will use.

It would also be good to have some more descriptive data about the whole corpus. This would normally be provided by the corpus header which, you will recall, at the moment has only the bare minimum provided when it was automatically created by Xaira Tools. You can edit the corpus header with any text editor, but an XML-aware one such as TEI-emacs is recommended.

- Close Xaira tools, and open the file `corpus_header.xml` with your chosen editor.
- Change the title of your corpus to something more specific than `varney`: (how about `Varney the Vampyre: an electronic corpus`?)
- Add text classification codes
- You could also add a brief description of the source of this corpus in the `<sourceDescription>`, such as `Found on the internet`; a `<revisionDesc>` with the date of this update; and any other header elements that appeal to you

Now re-open the Xaira Toolkit, and open your corpus_parameters.xml file again. Check that your changes to the index have been read, by choosing Bibliography from the Tools menu.

Finally, re-run the Indexer by selecting Indexer from the Tools menu as before.

Now try running the Xaira client against the new index. Repeat the same procedure as you used above to open the corpus, and then try some of the following commands:

1. Open the Word Query window and type in `horr` to see what horrific words appear in this text...
2. Choose one (how about `horrible`?) and then press the Query button.
3. If you look at the bottom right of the status bar as you step through the hits you will see that the references displayed change to indicate the chapter number and either the paragraph or the sentence number, depending on the choice you made when defining labels in the indexer.
4. Select one of the hits. To see the detail of the chapter it comes from, you can either right click with the mouse, and select "Source" from the submenu which appears; alternatively, press the "Source" button on the toolbar.
5. Now try out more of your favourite Xaira queries on this corpus!