Indexing with Xaira

Lou Burnard

Xaira – the XML version of Sara – is an XML Aware Indexing and Retrieval Architecture. This tutorial focuses on the Indexing aspects and shows how to build a Xaira searchable corpus.

Copy the file xaira-texts.zip from drive T: to your working directory D:documenti, and unpack it. It contains a folder called Samples containing a collection of sample texts for you to experiment with.

1 The bare minimum

We begin with the simplest case: a text file completely innocent of any markup. What can Xaira do with that?

In this first exercise we will make a tiny corpus, consisting of one chapter of some book called I Promessi Sposi which I found on the web. The file is in your Samples directory.

1. Open the Xaira Tools program by double clicking on its icon or locating it in the Program menu.
2. Select Index Wizard from the File Menu.
3. Give your corpus a name and a brief description.
4. Accept all the defaults in the next dialog and press Next (Xaira will create a new location for your corpus)
5. Use the Browse button to find the folder containing sample texts (`Samples`); press Next
6. The file you want to index is plain text, so select Plain Text and press Next
7. The wizard shows you a list of all the files in the Samples directory. You don't want to include them all, so press "Select/deselect files"
8. In the FileList dialog, select `sposi14.txt` and press the Select Button. Then press the OK button to return to the Wizard. Press Next.
9. In the language dialog, specify that this text is in Italian (it) and press Next
10. Press Index button. A DOS screen appears momentarily and then the message Indexer terminated with code 0 (OK)
11. Check the box labelled "Open Client" and then press Finish
12. The Xaira program opens: try out your favourite Xaira queries on this corpus!

2 Something a bit more ambitious

The Xaira wizard can also cope with text that is a bit more fully tagged. In this second exercise we will work on a chapter of Varney the Vampyre (probably the first vampire novel in English, since you ask) which has been automatically POS-tagged by CLAWS.

If Xaira Tools or Xaira is still running on your system, close them down before proceeding.

Proceed as before:

1. Open the Xaira Tools program by double clicking on its icon or locating it in the Program menu.
2. Select Index Wizard from the File Menu.
3. Give your corpus a name and a brief description.
4. Accept all the defaults in the next dialog and press Next (This is where you determine the location of your Xaira corpus)
5. Use the Browse button to find the Samples folder; press Next
6. The file you want to index this time is in XML, so select the XML radio button and press Next
7. Accept the defaults on the next screen (File Structure); press Next
8. Proceed as before, but this time, select the file called `varney-pos.xml`
9. After you have specified the language (en) for this text, Xaira will analyse the tagging it finds to decide which elements mark individual texts, sentences, and words. You need to confirm that it is right in each case, and also tell it how to refer to the individual texts, sentences, and words.
10. First, `Text delineation`. In this "corpus", each `<div>` marks a new text, and each div has an n attribute carrying its number. So select "n" from the right hand box, and press Next
11. In this corpus each `<s>` element also has an n attribute, so you can select n here again. Alternatively, you can select Auto-number, in which case the sentences will be numbered in sequence. Press Next.
12. Finally, confirm that Xaira is right in thinking that `<w>` is used to mark words in this corpus and press Next
13. This text has POS codes but no headwords (lemmas), so press Next on the next dialogue.
14. There is no bibliographic data in this file, so we press Next on the next dialogue too.
15. Press Go and index the corpus as before.
16. The Xaira program opens: try out your favourite Xaira queries on this corpus!

3 Beyond wizardry

The wizard can do a lot for you, but it's not omnipotent. The Xaira Tools program can help you build a Xaira system which can make more use of all the features in your markup.

In this exercise, we'll index a collection of publicity leaflets extracted from BNC text CFR. Like the rest of the BNC, this collection has detailed XML markup, down to the word level. We'll be exploring this markup in more detail tomorrow; for this exercise you need to know the following:

- Each `<div1>` element in the file contains a single pamphlet

- Each <s> element contains a sequence of words treated as a single unit by CLAWS
- Within <s> elements, <w> and <c> elements mark lexical items and punctuation marks respectively. Each of these elements carries a type attribute which supplies a part of speech code for it. The <w> elements also carry a <lemma> attribute which gives a headword corresponding with the word form it contains, for example the base form of an inflected verb.
- Each pamphlet has been given a descriptive code indicating its topic. This is supplied on the <div1> element as the value of its type attribute. The values used, and their meaning, are supplied in a <taxonomy> element in the TEI header prefixed to the collection.

3.1 Setting up the corpus

If Xaira Tools or Xaira is still running on your system, close them down before proceeding.

We begin by running through the same steps as the Wizard does when creating a new corpus.

1. Open the Xaira Tools program and select New from the File Menu.
2. Select Parameter File from the Tools menu
3. In the Parameters dialog, type `Leaflets` 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 `Samples/CFR` as the Root for this corpus.
4. Press the Defaults button. By default, Xaira expects to use four directories: one called Texts, one called Index, and one called Etc, and one called Usr: it will make them if necessary, so just press OK
5. 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 Texts folder (Samples/CFR/Texts) are to be indexed. Press the Generate button.
6. A list of the files in this directory appears. Use the Select or the Delete button as appropriate to ensure that only the file `leaflets.xml` appears in the lower window. Press OK
7. The file corpus_header.xml has been prepared for you already: it contains basic metadata for this collection of texts, but nothing specific to Xaira. Xaira will update it with any additional information it needs.
8. Select Refresh frm the Tools menu, and select Refresh All from the submenu which appears. menu. You should see the message `0 errors`. Press Save. Xaira now knows something about the markup in your document.

Now we will start telling Xaira how to use the markup in your documents.

1. Select Tag Usage from the Tools menu. A list of the elements 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. Let's try this now.

2. Select the caption 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 `Text in a display box` and press OK.

3. 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 document and the other one or more units within that document, such as a paragraph or sentence. In the whole BNC, for example, a reference such as CCC 123 identifies sentence number 123 in corpus text CCC. The unit is `sentence`, the document is `corpus text`.

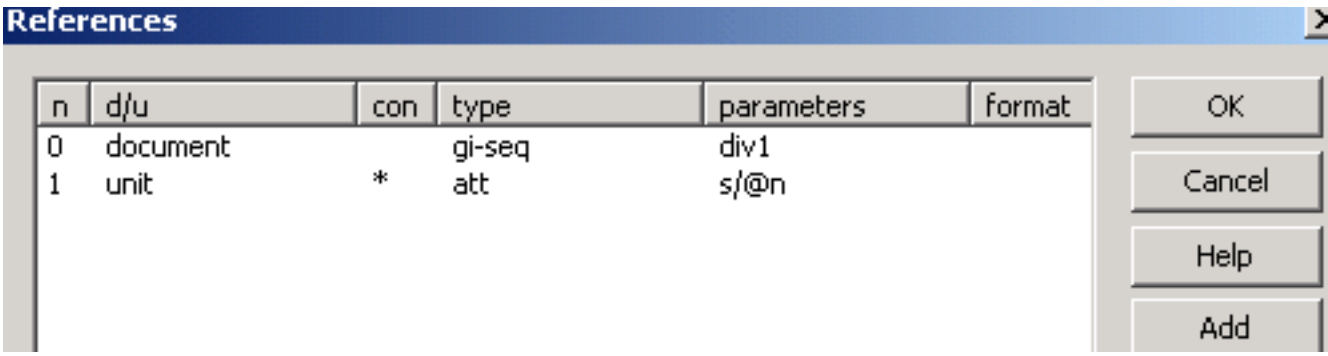
4. As you see, the default 'document' is a file, and the default 'unit' is a line-number. Since all of our text is in a single file, the current 'document' reference is not much use to us. To change it, first select the `0` to the left of the `document` line in the upper part of the References dialogue and press Edit.

5. For our leaflets it makes sense to treat each `<div1>` as a distinct document. In our tagging the leaflets do not have any explicit numbers, but Xaira can count them for us and then we can use the sequence number to identify them. Select the `Use gi sequence number` radio button, and then select `div1` from the available list of element names. Press OK.

6. The line number within the whole file is not a very useful way of identifying context within these pamphlets. Since we also have `<s>` elements, which carry numbers on their n attributes, we will use them.

7. First, select the `Use value of attribute on given element` radio button. Then select `div1` from the scrollable list of element names, and then select `n` from the list of available atributes by clicking on it. Finally, press OK.

8. The References dialog window should now look like this:



press the Save button if all is well. Press the help button if you would like to learn more about the other options available on this dialog.

3.2 Tokenizing and additional keys

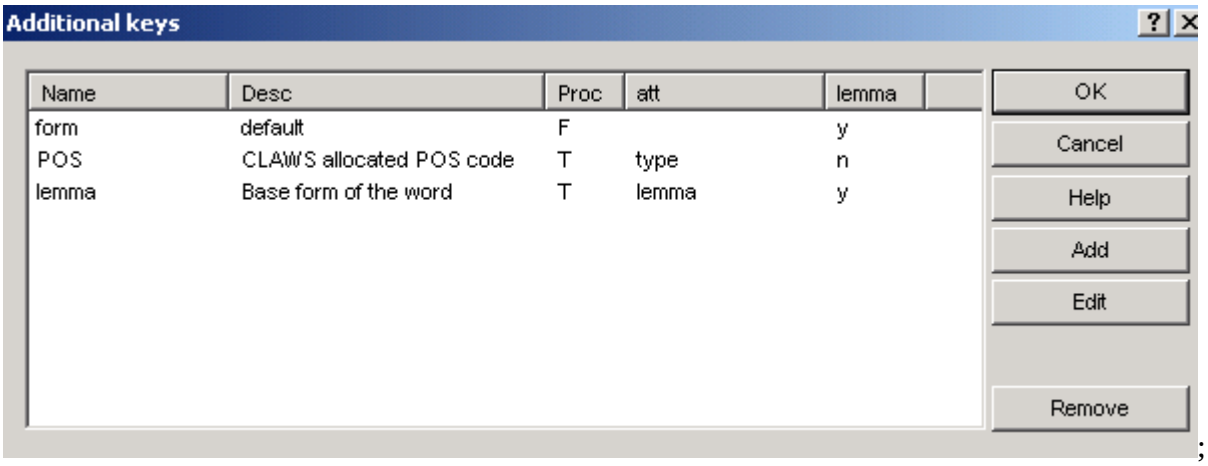
We now need to tell Xaira how to tokenize (split into lexical items) the text. Xaira can either simply split words at space characters, ignoring the markup, or it can use XML markup to indicate where words begin and end. In this corpus, we want it to use the `<w>` and `<c>` elements to mark word boundaries, irrespective of any spaces or other punctuation.

1. Select Special Tags from the Tools menu.

2. From the drop down list at the top, select Word Break
3. From the list at the bottom of the dialog box, click on w and c
4. Press OK

We also need to tell Xaira about the additional keys in our data. An additional key is a way of providing additional information about a word form, such as its part of speech, or headword. Xaira allows you to define as many additional keys as you need.

1. Select Additional Keys from the Tools menu
2. Click the Add button to define a new additional key
3. In the Additional Key dialogue, enter a name for the key ("pos") and a description ("CLAWS allocated POS code")
4. Select both `w` and `c` from the list of elements (hold down the CTRL key and click on each in turn). The `type` attribute is listed in the Attributes box, as this is the only attribute these two elements share. Press OK to add the value of this attribute as a key.
5. Click the Add button to add a new key
6. In the Additional Key dialogue, enter a name for the key ("lemma") and a description ("Base form of the word")
7. As before, select `w` from the list of Elements, and this time select `lemma` from the list of attributes. Check the box that says "lemma scheme" before pressing OK.
8. The Additional Keys dialog window should now look like this:



;
press the Save button if all is well. Press the help button if you would like to learn more about the other options available on this dialog.

Although we have defined an additional key called lemma, we have not yet defined a lemmatization scheme. A lemmatisation scheme is used by Xaira as an alternative way of indexing words in the corpus. For each word form to be indexed, Xaira needs to know one or more additional keys from which the lemma entry is to be constructed. In our case, this is simply the existing `lemma` additional key.

- Select the Lemma scheme command on the Tools menu
- A list of available additional keys appears. Select `lemma` and press OK

3.3 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 `<note>` elements in our 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 learn more about XSLT (the language used here) tomorrow.
- Change the line which currently reads `<xsl-copy-of select="teiHeader/SoureceStmt"/` to read `<xsl:ccopy-of select="note|div2/note"/>`: this means "copy the content of any `<note>` which is either directly contained by a document or by a `<div2>` within a document". Press OK and Save your changes
- 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.

3.4 Indexing the corpus

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.
- If you want to check that everything has worked correctly, use options on the Test menu. Two of these are particularly useful:

  Test Index – In the "View word" dialog, enter a word that interests you (try "vampire") and press the View button.
  - Click on the word that appears in the frequency list to see the different spellings corresponding with it.
  - Click on a word in the list of spellings to see the occurrences of it
  - Click on one of the occurrence lines, and see the context in which it appears.

  Test Bibliography – In the "Test Bibliography" dialog, the upper window shows the texts Xaira has identified in your corpus.
  - Click on the name of a text and the lower window shows the corresponding bibliographic description for it

- Finally, you need to create an xcorpus file for the xaira client. Select "xcorpus file" from the Tools menu.
- Close the Xaira Tools application.

3.5 Searching the corpus

You can run the Xaira client by simply double clicking on the "leaflets.xcorpus" file which the Tools command has created for you. Or you can start the client from the Program menu, choose Open from the file menu and navigate to the same file.

- Open the Word Query window and press the Lookup button without typing anything in. This will give you a frequency list of all the word forms in the collection in alpha order.
- Sort them into descending frequency order, by clicking on the frequency column. Do you see anything unusual in this list? (If you'd like to save it for further analysis, press the Save button)
- Choose a word (how about `you`?) and press the Query button at the bottom.
- 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 leaflet and sentence number.
- Select one of the hits. To see the detail of the leaflet it comes from, right click with the mouse and select "Source" from the submenu which appears; alternatively, press the "Source" button on the toolbar.
- Why not try out more of your favourite Xaira queries on this corpus?

3.6 Setting up a partition

A partition is a way of dividing up a corpus which, like this one, contains a number of different kinds of text. You can define partitions in many ways, ranging from the purely impressionistic to the rigorously formal. Once a partition has been defined for Xaira, you can analyse the lexicon of the different kinds of text (classes) making up the partition. In this exercise, we will define a partition on the basis of the values of the type attribute supplied for each leaflet in our collection. These values were added for the purpose of this exercise and have no particularly significance beyond that!

1. First open the texts window. A list of the documents defined in our corpus appears. Choose Column Control from the Texts menu.
2. The upper part of the Text Windows Columns dialogue shows you all the available elements in your corpus; the lower part shows you the column headings in the Text display (other than Text and Class). This dialogue allows you to add new columns to this display. The column contents will be derived from either attribute values or element content, which you select from the list in the upper part.
3. Choose `div1` from the scrollable lift in the upper part of the dialogue and click the Attribute radio button. From the list of attributes displayed, select "type" and press the Add button. Close the window. In the Text Window, each document now has a classification code.
4. We will now define a new partition, based on the column we have just added. Click on the New Partition button, to open the New partition dialogue. You will need to supply a name and description for the partition: we suggest `domain` and `Arbitrary topic assignation` respectively.

5. Check the second radio button, which is labelled "Create a partition based on values in a column". Select "domain" from the list of available columns and press the OK button.

6. A new column opens in the text window: each row now contains, by default, the values supplied in the column you nominated when you defined this partition. You can reclassify them if you like.

7. Choose Partition Properties from the Texts menu, and you will see that each of your classes has been allocated a colour. You can change the colour by selecting the class name and pressing the Colour button. You can also add (but not delete) a class in this dialogue.

8. Now choose Activate Class from the Texts menu. A window appears showing the available class codes: choose one (say "M-KW"). In the status bar, you will see that the corpus name is now suffixed by the class code ("Samples:M-KW"): this indicates that only the M:KW texts in the Samples corpus are now being searched. Check this by doing a search for any word (say "you"): the results will be taken only from texts with this classification.

9. The second window on the lower button bar now also contains a list of available classifications. Choose a different classification (say "IN-H") and repeat your search. A second query window now opens, with results taken only from the "IN-H" texts in the corpus. (You may find it convenient to place the two query result windows side by side by choosing Tile from the Windows menu). You can now compare the usage patterns of the word "you" in texts classified differently in your corpus.

10. Finally, we will activate the whole partition created by the "domain" classification scheme. Choose "Activate partition" from the Text menu. A list of available partitions appears: select "domain" and press OK. Repeat your search for "you", and you will now see a breakdown of how the hits for this query are distributed across different kinds of leaflet.

11. Select Analysis from the Query menu. A new window opens showing at the top various statistical properties of your partition, and at the bottom a graphic display, as either a bar or a pie chart.

The columns show for each row

hits The number of words matching your query found in texts allocated to the specified class

words The size in words of the texts allocated to the specified class

% hits as a percentage of words

Hit Texts the number of texts allocated to the specified class which contain at least one occurrence of words matching your query

Texts the number of texts allocated to the specified class

If you check the box labelled ‘Measure size of hits’, the "Hits" column (and consequently the "%") column changes to indicate the size in words of all the texts containing at least one hit.

You can save these statistics in a file by clicking the Listing button. You can also copy the graphic to the Windows clipboard, by pressing the Copy button.