Exercise: Basic XAIRA Use

TEI@Oxford *2009-07*

XAIRA (XML Aware Indexing and Retrieval Architecture) is a complex tool which can be used to explore any kind of XML corpus (not necessarily TEI!) in various ways. This exercise uses two parts of it only: the Windows tool kit, and the Windows client. There are other components for other platforms or environments, e.g. for developing web-hosted applications using PHP or Java. See `http://www.xaira.org` for details.

XAIRA is an open source project developed at OUCS, initially for use with the British National Corpus (`http://www.natcorp.ox.ac.uk`); for more detailed instructions (including screen-shots) about how to use XAIRA with the BNC see `http://www.natcorp.ox.ac.uk/tools/bncXml_search.xml`; for more information about the XAIRA client, see `http://www.oucs.ox.ac.uk/rts/xaira/Doc/refman.xml` (This online manual is also available as a helpfile within the XAIRA client).

1 Getting started

Before you can use XAIRA with your corpus you need to install the software and index your corpus. If the XAIRA client is not already installed on your desktop, you can download it from `http://www.oucs.ox.ac.uk/rts/xaira/Download/xaira124.zip` Click on this file to unzip it, and then click on the resulting `msi` file to install it. Accept all the defaults.

XAIRA will index a corpus which has almost any level of XML markup, including none at all. For this exercise, we'll use a version of the Punch files we have prepared which includes some automatically-generated linguistic markup, and (just to show we can) another version in which there is no markup at all.

The files you need are all available in the teidata folder: just copy the zip file p4x.zip from it to your Desktop, and click on it to unzip it. You should get a folder called p4x (Punch For Xaira) containing two subfolders, Plain and XML.

XAIRA has two components: the XAIRA client, which is used to search a XAIRA database, and Xaira-tools, which is used to build one. Start by clicking on the icon for Xaira Tools, which should be on your START menu when you have installed the system.

1.1 Indexing a corpus with no mark-up

1. Choose 'Index Wizard' from the File menu

2. The Corpus Name window opens. Supply a name for your corpus (preferably a simple name with no spaces or punctuation marks in it) such as Punch-Plain ; you can also supply a more detailed description of your corpus on this screen, if you wish. Click the NEXT button to continue.

3. The Corpus Root window opens. XAIRA has created a folder of this name in which it will store all the files created by the indexer. By default this folder is placed inside the folder My Corpora, inside Documents and Settings. Click the NEXT button to continue.

4. The Texts window opens. You need to tell XAIRA where to find the texts for your corpus. Click on BROWSE and navigate to the p4x folder where your files were unpacked. Choose Plain. Click the NEXT button to continue.

5. The Markup window appears. You need to tell XAIRA about the markup of your texts. These files have no markup at all: so check that the radio button "Plain Text" is selected. Click the NEXT button to continue.

6. The File list window opens, and shows a list of all the files available in this folder so that you can pick and choose. We will process them all. Click the NEXT button to continue.

7. The Character encoding window appears. XAIRA will try to identify the character encoding (e.g. Unicode, ISO-88591, Windows...) of your files automatically. Click the NEXT button to continue.

8. The Reading files window appears. XAIRA needs to check that your files are all valid. Click on GO. Behind the scenes, XAIRA has added a very minimal amount of XML tagging to your documents and is now checking that. Click the NEXT button to continue.

9. The Language window opens. Choose the appropriate language (en) from the list. Click the NEXT button to continue.

10. The Indexing window appears. When you click on the INDEX button, XAIRA will get to work, processing all your texts, and making a number of index files. It will also create (in the etc directory) a log recording each step of the process, which can be useful if something goes wrong.

11. Select the "View corpus in Xaira client" tick box. Click the FINISH button to close the Wizard.

Your corpus is ready for use! What can we find out with this version of the Punch corpus?

- you can search for word forms (look for the word *german*) or word patterns (type `...` to find all three letter words in the corpus; note that the pattern check box needs to be checked for this to work!)
- you can identify where each occurrence of a word is in the text, but XAIRA will only show you a single line of context
- the location for each occurrence is given by the name of the file and the line number within the file, displayed at bottom right of the screen.

1.2 Indexing a richly-encoded corpus

In these files we have marked up the word boundaries using the `<w>` XML element. As well as allowing us to tokenize the text explicitly, this also allows us to annotate each word with a part of speech and a root form, using attributes type and lemma respectively. This was done using a freely available tagging program called tree tagger developed at the University of Stuttgart; tree tagger also enables us to identify sentence divisions automatically in the text, with moderately successful results. These are useful for delimiting searches and giving us rather more accurate reference information, as we will see.

1. Start up the XAIRA TOOLS application again (if it is still open, you need to close it first)

2. Select 'Index Wizard' command from the File menu

3. The `Corpus Name` dialog opens. Enter a name for your corpus ("Punch" for example), and a description of the corpus if you like. Press NEXT to continue.

4. The Corpus Root dialog opens. Press NEXT to continue.

5. The Texts dialog opens. Press BROWSE and navigate to the folder that you unzipped the p4x.zip file into. This time select the folder named XML. Press NEXT to continue.

6. The Markup dialog opens. Select the radio button "XML".

7. The File Structure dialog opens. Choose Model 1 and press NEXT to continue with the default options.

8. The File list dialog opens. You could select one or two files only here, but by default we will index the whole lot. Press NEXT to continue.

9. The Reading files dialog opens. Press GO to validate the files: there should be no errors. Press NEXT to continue.

10. The Language window opens. Choose the appropriate language (en) from the list. Click the NEXT button to continue.

11. The next few questions determine how lines of context taken from corpus texts are to be identified or referenced by the client. For each context in a concordance window, XAIRA will construct a reference identifying its location. This has at least two parts: the text identifier and one or more unit identifiers.

12. On the Text Delineation dialog, check the box that says "Just use file names". Press NEXT to continue.

13. On the Unit Delineation dialog select the element `s` from the scrolling list on the left, and select the `Auto-number` from the window on the right. Press NEXT to continue.

14. On the Word Delineation dialogue, select the `<w>` element and press NEXT.

15. The Additional Keys dialogue opens: the default keys identified by the wizard are the attributes type and lemma. Press NEXT to continue.

16. The Bibliography dialogue appears; we haven't prepared this for you, so press NEXT to continue.

17. The Indexing dialog appears as before. For this corpus, we need to take one more step which the Wizard cannot help us with. Do *not* press Index yet! Instead, press Cancel to close the wizard.

18. The Corpus Wizard cannot do everything. The XAIRA Tools utility allows you to fine tune almost all aspects of your corpus indexing. We will demonstrate this by defining what XAIRA calls a *lemma scheme* for your corpus. Proceed as follows:

    - In XAIRA Tools, select 'LemmaSchemes' from the Tools menu. The Lemma Schemes dialog opens.
    - Press Add to create a new lemma scheme: a second Lemma scheme dialog box opens.
    - In the Name box, enter "TT" for the lemmatization scheme
    - In the Gloss box, enter "Tree Tagger", since this is the agency solely responsible for this lemmatization scheme
    - Choose lemma from the list of available additional keys and press the Add button to add it to the lemma scheme
    - Press OK to return to the first Lemma Schemes dialog, and press OK again.
    - One further step is necessary to cope with the way the OUCS lecture room machines are set up. Select Indexer/Options to open the indexer options dialog. Uncheck the box that says `Calculate location buffer size automatically`. Enter a suitable size for this buffer (we suggest 3000000).
    - You are now ready to index your texts. Select Indexer/Run from the Tools menu.

19. When indexing is complete, start the XAIRA client, select Open from the File menu, and navigate to the xcorpus file created in My Corpora\Punch to open it.

2 Exploring the corpus with XAIRA

2.1 Basic search

Use the search box on the XAIRA toolbar to search for **german**. Download just the first 100 solutions, or a random selection.

2.2 Display of solutions

1. Toggle between Page mode and Line mode (Use the Page/line mode button)
2. Display more context. (Use Scope list and change to 2)
3. Display the XML (Use the Format list , select XML) .
4. You can also make your own stylesheet to display things in different ways.
5. Sort the lines on the word immediately following the search word (Use Sort button. Key: One; Sort Key: Right, Ascending, Span 1. Tick 'Use text', Collating: Primary).

2.3 Collocation

The collocation feature allows you to see what co-occurs with a particular search term.

1. Use the Collocation button to retrieve the collocations of your search word. Tick the 'Controls' box (bottom of collocation window). (Use the Collocations function. Untick 'Downloads only'. Set window 0L 1R)
2. Select the 'Window' tab and set the span to Left 3 Right 3. Click 'calculate'. Explore the result.
3. Click the colligation tab, tick 'colligation' and select TYPE. Explore result.
4. Try some other words (such as "English" or "French") to see if they have similar collocation patterns.

2.4 More search options

XAIRA offers a wide range of search options. You can make the same search in different ways. You can also combine different kinds of searches.

Include part-of-speech and/or lemma information for search term Use the Word Query command: tick boxes 'Controls' and 'Unique forms' at bottom of window) . Type 'round' in the search box and click 'Lookup'. Click on the word 'round' in the result list and see what is displayed in the bottom window (= different word-class tags available with 'round'). Look for lemma information: Select the 'Lemmata' tab (in right-hand part of window). Choose 'tt' from the drop-down menu. What happens in the left-hand part of the window? Click on a word in the top window and explore the result in the bottom one.

Search by regular-expression You can use regular expressions in two places:

- Find a list of all four-letter words in the corpus using the regular-expression tickbox in Word Query
- Find all occurrences of four-letter words in the corpus using the Pattern Query)

Search by word-class information Use the Addkey query. Tick the 'Any' box. Click 'Refresh' to display all available part of speech tags. Select the one(s) you want to search for and click OK (select more than one by pressing the CTRL-key on your keyboard while you click).

Search XML markup Use XML Query to do markup-sensitive queries. For example:

- Look for `<persName>` elements. Sort them by the right 3 words.
- Find what values have been assigned to the type attribute on `<div>`

Combine search options Use the Query Builder to find ...

- sentences containing three consecutive adjectives
- occurrences of 'dog' and 'canadian' within 10 words of each other