Wmatrix: a web-based corpus processing environment.

Paul Rayson (paul@comp.lancs.ac.uk)
Computing Department, Lancaster University, UK.

In this software demonstration, I will introduce Wmatrix, a web-based environment which allows researchers at Lancaster to have local and remote access to some of UCREL's corpus annotation and retrieval tools. The web browser provides a much simpler interface to these tools than via the UNIX command line. All processing is done on the remote web server so users gain access from any platform that provides a browser such as Netscape or Internet Explorer. Tools available in Wmatrix include CLAWS (part-of-speech tagger), SEMTAG (word-sense tagger) and LEMMINGS (a lemmatiser). Wmatrix also provides production of frequency lists, statistical comparison of those lists, and KWIC concordances.

Wmatrix was built during REVERE (Rayson et al, 2000), a UK funded project investigating the extraction of information from software engineering documents. One of the aims of the project was to investigate the use of NLP tools to aid software engineers in their understanding of a software system. The information on the software system is contained in existing documentation or transcripts and reports from ethnographic studies of the system being used. We built a web-based information extraction environment by locating various UCREL NLP tools on a web server and by providing the Wmatrix interface to those tools. The output of the tools can be presented in a web browser from different viewpoints depending on the role taken by the user of the system, but this demonstration will be from the corpus linguist viewpoint. This presents the traditional model of submitting raw data to Wmatrix, passing it through the corpus annotation tools and then using concordances to view the results.

A user of Wmatrix begins by uploading their corpus to the web server via a web browser such as Netscape Navigator or Microsoft Internet Explorer. The first corpus annotation tool applied to the text is the hybrid part-of-speech tagger, CLAWS (Garside and Smith, 1997) which assigns a part-of-speech tag to every word in running text with about 97% accuracy. A second layer of annotation is applied by SEMTAG, a semantic tagger (Rayson and Wilson, 1996). This tool assigns a semantic field tag to every word in the text with about 92% accuracy. The resulting annotated files are presented to the user in a workarea and Wmatrix prepares word, POS and semantic tag frequency lists. These can be downloaded but can also be browsed using the web browser application. The user can select a word or tag from the lists and see a standard key word in context concordance for that item. This is prepared on the fly from the corpus on the web server.

Users are guided towards interesting words or tags to investigate further by comparing frequency lists from their corpora to standard textual norms provided by frequency lists produced from the British National Corpus for example.

Each user of Wmatrix has their own set of workareas containing corpora that they have processed. Wmatrix is designed to cope with corpora up to several million words in size, but retrieval would be less interactive with larger corpora. A web based interface for the Stuttgart Corpus WorkBench is available. The Corpus WorkBench (Christ, 1994) pre-indexes the text and is consequently much faster at providing concordances for large corpora. I am currently working on integrating this into Wmatrix so that texts can be automatically indexed for CQP queries.

Acknowledgements

The REVERE project (REVerse Engineering of Requirements) was supported under the EPSRC Systems Engineering for Business Process Change (SEBPC) programme, project number GR/MO4846. Further details can be found at: http://www.comp.lancs.ac.uk/computing/research/cseg/projects/revere/

The web interface to IMS CWB was provided by Tomaz Erjavec as used in the Slovene corpus tool at: http://nl2.ijs.si/corpus/

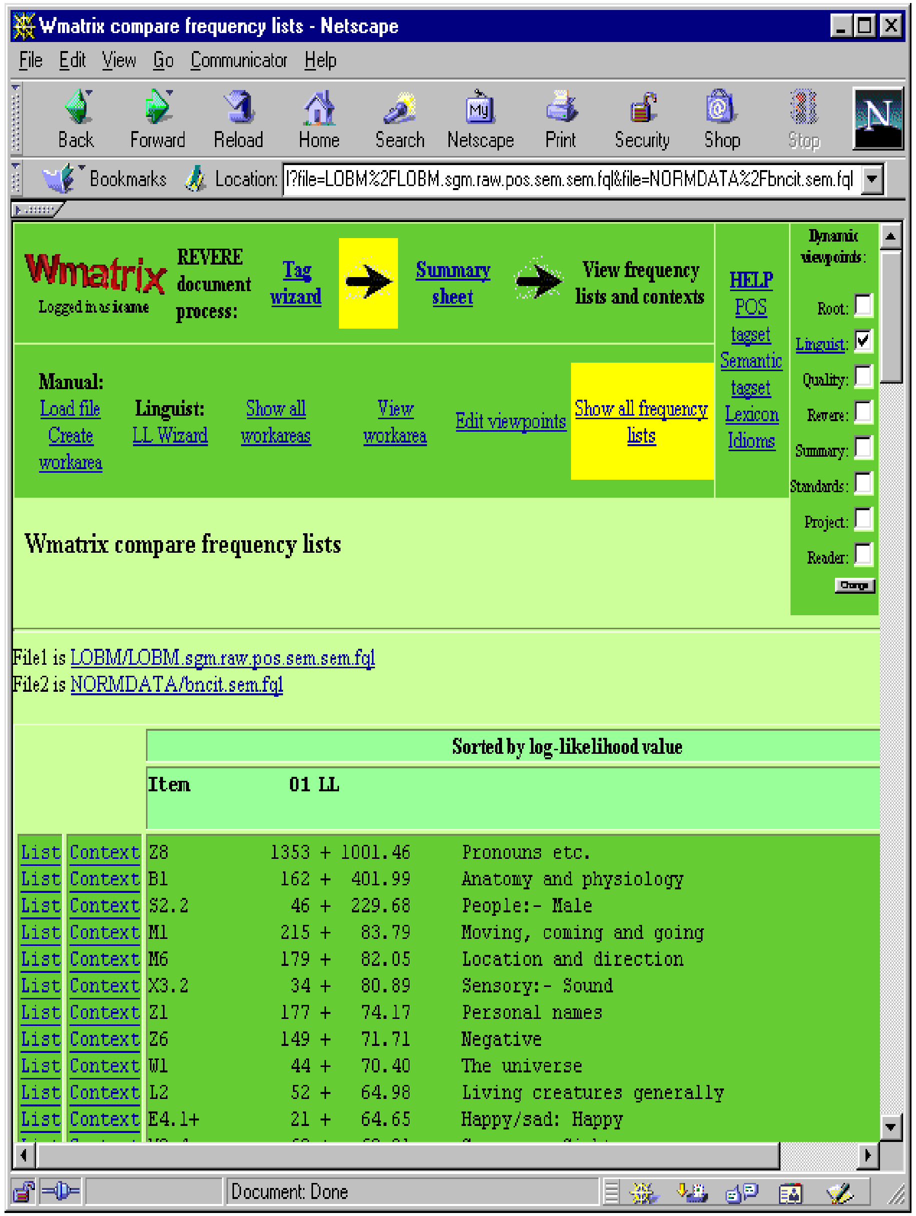
References

Christ, O. (1994). A modular and flexible architecture for an integrated corpus query system. Proceedings of COMPLEX'94: 3rd Conference on Computational Lexicography and Text Research (July 7-10 1994). Budapest, Hungary. pp23-32.

Garside, R., Smith, N. (1997). A Hybrid Grammatical Tagger: CLAWS4, in Garside, R., Leech, G., and McEnery, A. (Eds.) *Corpus Annotation: Linguistic Information from Computer Text Corpora*, London: Longman, pp. 102 - 121.

Rayson, P., and Wilson, A. (1996). The ACAMRIT semantic tagging system: progress report. In L. J. Evett, and T. G. Rose (eds) *Language Engineering for Document Analysis and Recognition, LEDAR, AISB96 Workshop proceedings*, pp 13-20. Brighton, England.

Rayson, P., Garside, R., and Sawyer, P. (2000). Assisting requirements engineering with semantic document analysis. In Proceedings of *Content-based multimedia information access RIAO 2000 International Conference*, College de France, Paris, France, April 12-14, 2000. C.I.D., Paris, pp. 1363 - 1371.

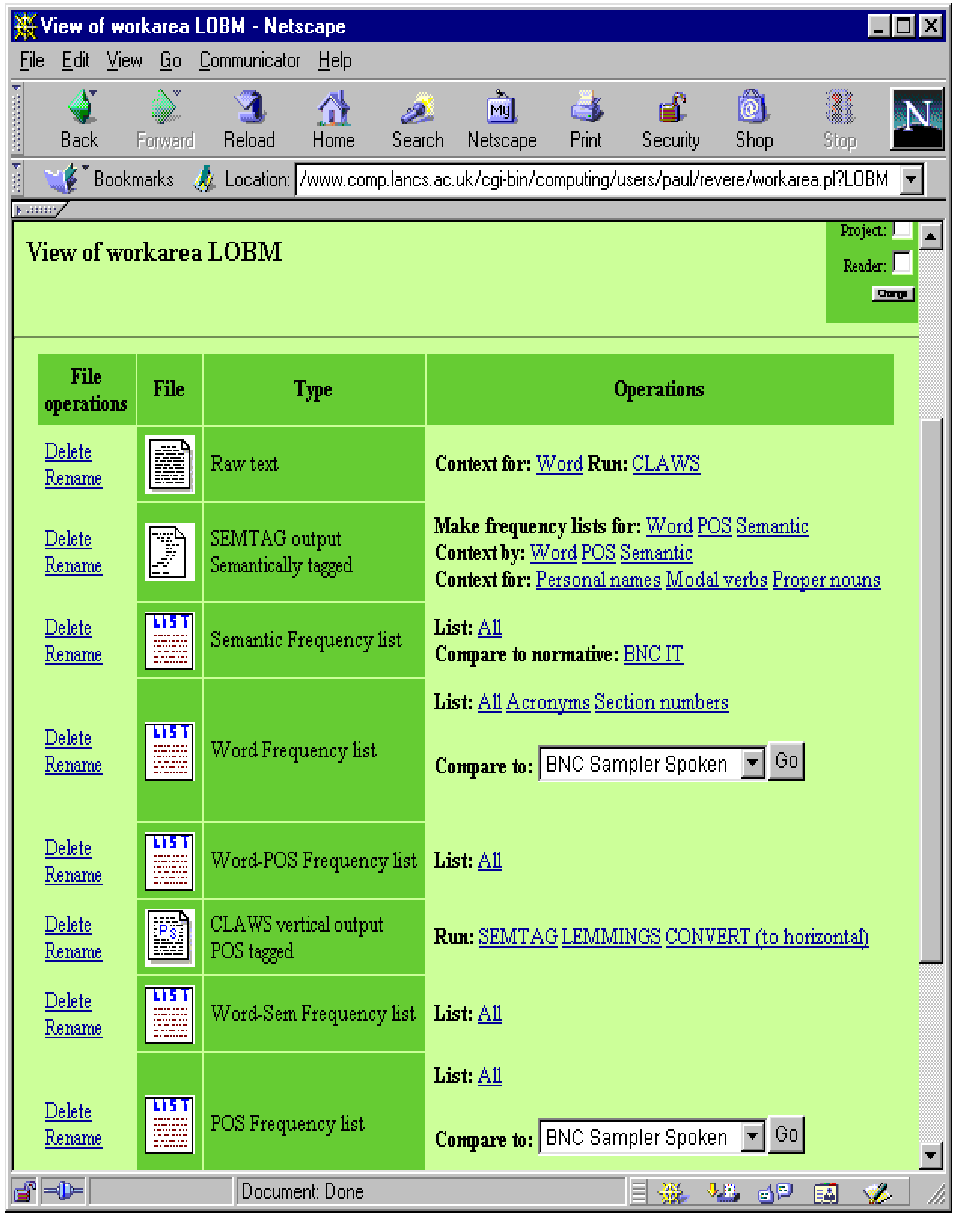

Wmatrix compare frequency lists - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Bookmarks Location: l?file=LOBM%2FLOBM.sgm.raw.pos.sem.sem.fql&file=NORMDATA%2Fbncit.sem.fql
Wmatrix
Logged in as icame
REVERE document process: Tag wizard → Summary sheet → View frequency lists and contexts
HELP POS tagset Semantic tagset Lexicon Idioms
Manual: Load file Create workarea
Linguist: LL Wizard
Show all workareas
View workarea
Edit viewpoints
Show all frequency lists
Dynamic viewpoints:
Root: □
Linguist: ☑
Quality: □
Revere: □
Summary: □
Standards: □
Project: □
Reader: □
Change
Document: Done

Wmatrix compare frequency lists

File1 is LOBM/LOBM.sgm.raw.pos.sem.sem.fql
File2 is NORMDATA/bncit.sem.fql

Sorted by log-likelihood value

```
             Item          01 LL
List Context Z8          1353 + 1001.46    Pronouns etc.
List Context B1           162 +  401.99    Anatomy and physiology
List Context S2.2          46 +  229.68    People:- Male
List Context M1           215 +   83.79    Moving, coming and going
List Context M6           179 +   82.05    Location and direction
List Context X3.2          34 +   80.89    Sensory:- Sound
List Context Z1           177 +   74.17    Personal names
List Context Z6           149 +   71.71    Negative
List Context W1            44 +   70.40    The universe
List Context L2            52 +   64.98    Living creatures generally
List Context E4.1+         21 +   64.65    Happy/sad: Happy
```


View of workarea LOBM - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Bookmarks Location: /www.comp.lancs.ac.uk/cgi-bin/computing/users/paul/revere/workarea.pl?LOBM
Project: □
Reader: □
Change
Document: Done

View of workarea LOBM

File operations	File	Type	Operations
Delete Rename		Raw text	**Context for:** Word **Run:** CLAWS
Delete Rename		SEMTAG output Semantically tagged	**Make frequency lists for:** Word POS Semantic **Context by:** Word POS Semantic **Context for:** Personal names Modal verbs Proper nouns
Delete Rename	LIST	Semantic Frequency list	**List:** All **Compare to normative:** BNC IT
Delete Rename	LIST	Word Frequency list	**List:** All Acronyms Section numbers **Compare to:** BNC Sampler Spoken ▼ Go
Delete Rename	LIST	Word-POS Frequency list	**List:** All
Delete Rename	PS	CLAWS vertical output POS tagged	**Run:** SEMTAG LEMMINGS CONVERT (to horizontal)
Delete Rename	LIST	Word-Sem Frequency list	**List:** All
Delete Rename	LIST	POS Frequency list	**List:** All **Compare to:** BNC Sampler Spoken ▼ Go