Burnard, Lou & Ylva Berglund Prytz: Exploring BNC XML Edition with Xaira

This software demonstration will introduce the Xaira program and illustrate how it can be used with a large, richly annotated corpus, the *BNC XML Edition*. Among the new features available to users of Xaira and BNC XML Edition are options to:

- search for tags only ('all –ing-forms of verbs', 'preposition + that', etc.),
- quickly locate and search subcorpora defined by existing text categories (genre, written/spoken, year of publication, target audience, etc.),
- define searchable subcorpora according to your own categorization,
- display search result as graphs,
- see distribution across text categories (existing or user-defined),
- browse a text without seeing the annotation.

Xaira (XML Aware Indexing and Retrieval Architecture) is based on SARA, the text searching software originally developed at Oxford University Computing Service for use with the British National Corpus. Xaira can be used on any corpus of well-formed XML documents in order to search and display text in any language (after processing). It is distributed under a GNU General Public License (GPL) and can be downloaded for free from the Xaira SourceForge site (http://xaira.sourceforge.net/)

More information:
BNC webpage: http://www.natcorp.ox.ac.uk/
Xaira webpage: http://www.oucs.ox.ac.uk/rts/xaira/
Xaira Reference Guide: http://www.oucs.ox.ac.uk/rts/xaira/Doc/refman.xml
Xaira SourceForge site: http://xaira.sourceforge.net/

Granger, Sylviane & Fanny Meunier: The International Corpus of Learner English - version 2: not just more, simply better!

This software demonstration will present the second version of the *International Corpus of Learner English* (Granger et al forthcoming). The enlarged database and enhanced query system make it a powerful versatile resource which will allow researchers to do full justice to variability in interlanguage.

ICLE 2 offers access to a larger number of subcorpora of advanced written learner language. Beside the 11 subcorpora included in the first version of *ICLE* (Granger et al 2002, Granger 2003), it contains data from five new mother tongue backgrounds, three of which outside Europe (Chinese, Japanese, Norwegian, Tswana, Turkish). In addition, the new interface contains a built-in concordancer which will allow researchers not only to draw up concordances of the words/phrases they are interested in, but also to get a breakdown of the search strings in terms of the many demographic and task variables recorded in the *ICLE* database (mother tongue of the speakers, age, number of years of English, time spent in an English-speaking country, text length, topic, text type, etc.). The system is based on the UNITEX corpus processing tool (Paumier 2002) which allows for sophisticated searches at various levels, e.g. word-form, lemma and part-of-speech. An important feature of the way we integrated UNITEX is that built-in dictionaries were not used as the software does not provide routines for disambiguation. The texts were POS-tagged with CLAWS C7 (cf.http://www.comp.lancs.ac.uk/ucrel/claws/) and the dictionaries for *ICLE 2* were compiled from the POS-tagged output.

The software demo will present the new options included in *ICLE 2* together with concrete illustrations of the new facilities.

References

Granger, S. (2003) The International Corpus of Learner English: A New Resource for Foreign Language Learning and Teaching and Second Language Acquisition Research. *TESOL Quarterly* 37(3), 538-546.

Granger, S., Dagneaux, E. & Meunier, F. (2002) *The International Corpus of Learner English. Version 1.1*. Handbook & CD-ROM. Presses universitaires de Louvain: Louvain-la-Neuve. Available from http://www.i6doc.com

Granger, S., Dagneaux, E., Meunier, F. & Paquot, M. (forthcoming) *The International Corpus of Learner English. Version 2.* Handbook & CD-ROM. Presses universitaires de Louvain: Louvain-la-Neuve.

Paumier, S. 2002. *Manuel d'utilisation d'Unitex*. Downloadable from http://ww-igm.univ-mlv.fr/~unitex/)

Kehoe, Andrew & Matt Gee: The WebCorp Linguist's Search Engine

This software demonstration will present the WebCorp Linguist's Search Engine (http://www.webcorp.org.uk), a tool which allows linguists to search the web as a corpus on a vast scale. Building upon previous work (Renouf et al, 2007; Kehoe & Gee, forthcoming), we show how the Search Engine has been designed to overcome the limitations of our existing WebCorp system by bypassing commercial search engines and building web corpora of known size and composition**.** We introduce the search interface and demonstrate new functionality, including:

- pattern matching and regular expression search (including wildcards)
- grammatical search (part-of-speech tags only or combined lexical/POS search)
- language specification
- textual domain specification
- sentence and paragraph boundary detection
- sentence position selection (within single sentence, sentence initial, sentence final)
- statistical collocation (external and phrase internal)
- diachronic search and graphical plotting of results
- sorting of output (by position, date)
- query refinement and concordance filtering
- web page caching

References

Kehoe, A. & M. Gee (forthcoming) 'New corpora from the web: making web text more 'text-like'' in *Proceedings of ICAME 2006, Helsinki*.

Renouf, A., A. Kehoe & J. Banerjee (2007) 'WebCorp: an integrated system for web text search' in C. Nesselhauf, M. Hundt & C. Biewer (eds.), *Corpus Linguistics and the Web.* Amsterdam: Rodopi.

Rayson, Paul: Word clouds and spelling variation

In this software demonstration, subtitled "Corpus analysis and annotation reassessed", I will demonstrate two separate tools Wmatrix and VARD. This will illustrate practical problems of applying analysis and annotation techniques, developed for modern corpora, to historical datasets.

The first tool, Wmatrix[1] is a web-based tool which permits data to be annotated automatically for parts-of-speech using CLAWS (Garside and Smith, 1997) and semantic fields using USAS (Rayson et al, 2004). Wmatrix was previously demonstrated at ICAME 2001 in Louvain and shown to extend the key words technique (Scott, 1997) to key domains (Rayson, 2005). In this presentation I will focus on the new "Word Clouds" visualisation of the key words. The word clouds display was inspired by the tag clouds of folksonomies as seen, for example, on the online photo sharing tool Flickr[2] and the social bookmarking site del.icio.us[3].

The second tool, VARD, allows the user of an historical corpus to identify spelling variants and link them to modern equivalents. By inserting modern equivalents in the corpus alongside spelling variants, standard corpus analysis and annotation techniques can be applied more easily and with greater accuracy. Without this pre-processing phase, techniques such as frequency profiling, part-of-speech tagging, collocations, n-grams and key words are much less robust than we would like them to be (Rayson et al, 2006).

[1] http://www.comp.lancs.ac.uk/ucrel/wmatrix/
[2] http://www.flickr.com/photos/tags/
[3] http://del.icio.us/tag/

References

Garside, R., and 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*. Longman, London, pp. 102-121.

Rayson, P., Archer, D., Piao, S. L., McEnery, T. (2004). The UCREL semantic analysis system. In *proceedings of the workshop on Beyond Named Entity Recognition Semantic labelling for NLP tasks in association with 4th International Conference on Language Resources and Evaluation (LREC 2004)*, 25th May 2004, Lisbon, Portugal, pp. 7-12.

Rayson, P., Archer, D., Baron, A. and Smith, N. (2006). Tagging historical corpora - the problem of spelling variation. In *proceedings of Digital Historical Corpora, Dagstuhl-Seminar 06491, International Conference and Research Center for Computer Science, Schloss Dagstuhl*, Wadern, Germany, December 3rd-8th 2006.

Rayson, P. (2005). Keywords are not enough. Presented at the *joint 26th ICAME and 6th AAACL conference*, Ann Arbor, Michigan May 12-15, 2005.

Scott, M. (1997) PC Analysis of Key Words -- and Key Key Words, *System*, Vol. 25, No. 1, pp. 1-13.

Wallis, Sean: ICECUP: The Next Generation?

We are developing a computer program for linguists to carry out complex, statistically sound experiments on a large, grammatically analysed corpus.

Currently this is time consuming, error prone and difficult, and only simple experiments are feasible. The aim of the project is to integrate and entire experimental cycle (definition, sampling, analysis and evaluation) into the same software suite, ICECUP 9Nelson, Wallis and Aarts 2002).

This has three immediate benefits. Previously difficult tasks become automatic. New procedures – including automatic enumeration of discrete and numeric variables, evaluation of case interaction (Wallis 2007), and the analysis of several variables in combination – become possible. Finally, linguists can comprehend their results by reference to cases in the corpus.

The project builds on the familiar ICECUP 3.1 software. Using ICECUP as a basis, researchers can formalise the simple exploration of a dedicated phenomenon by developing a series of experiments to investigate it – or 'informally' explore a phenomenon revealed by experiment.

More information on the project is at www.ucl.ac.uk/english-usage/projects/next-gen.

References

Nelson, G., Wallis, S.A. and Aarts, N. (2002). *Exploring Natural Language: Working with the British Component of the International Corpus of English* (Varieties of English around the World series), Amsterdam: Benjamins.

Wallis, S.A. (forthcoming, 2007). Searching treebanks and other structured corpora. Chapter 36 in Lüdeling, A. & Kytö, M (ed.) *Corpus Linguistics: An International Handbook.* Handbücher zur Sprache und Kommunikationswissenschaft series. Berlin: Mouton de Gruyter.