Xaira: software for language analysis

This paper describes XAIRA (pronounced like "Sarah" with a bit more voicing at the beginning), a software environment developed specifically for corpus-based analysis over the last decade at Oxford University Computing Services (OUCS). XAIRA supports a range of basic concordancing, indexing, and analysis requirements, and can be used to search through any XML corpus, in any language, large or small, with rich or only minimal XML encoding. It is written in C++, and is distributed under an open source GNU Public Licence.

The origins of the Xaira system go back to the final stages of the original British National Corpus project in the mid 1990s, when the various partners in the project agreed that the 100 million word corpus they were about to deliver to the world really needed to be provided along with some kind of software search engine. Some of the basic design decisions involved in developing that search engine have proved their value and are retained in XAIRA while others have been overtaken by events.

XAIRA, like SARA its predecessor, is a retrieval only system: it operates on a richly encoded text, but contains no enriching procedures itself. Both systems operate according to the *client-server* paradigm, in which the processes of providing access to the corpus, and interacting with the corpus user are carried out by different components, often not even on the same computer. But where SARA was designed to work with just one corpus using its own well-defined format, XAIRA is designed to operate with any corpus in any XML format; and where SARA was a fixed body of proprietary code, XAIRA is a modular system of components written to conform to an open *Applications Program Interface (API)* and derived from an explicit Object Model.

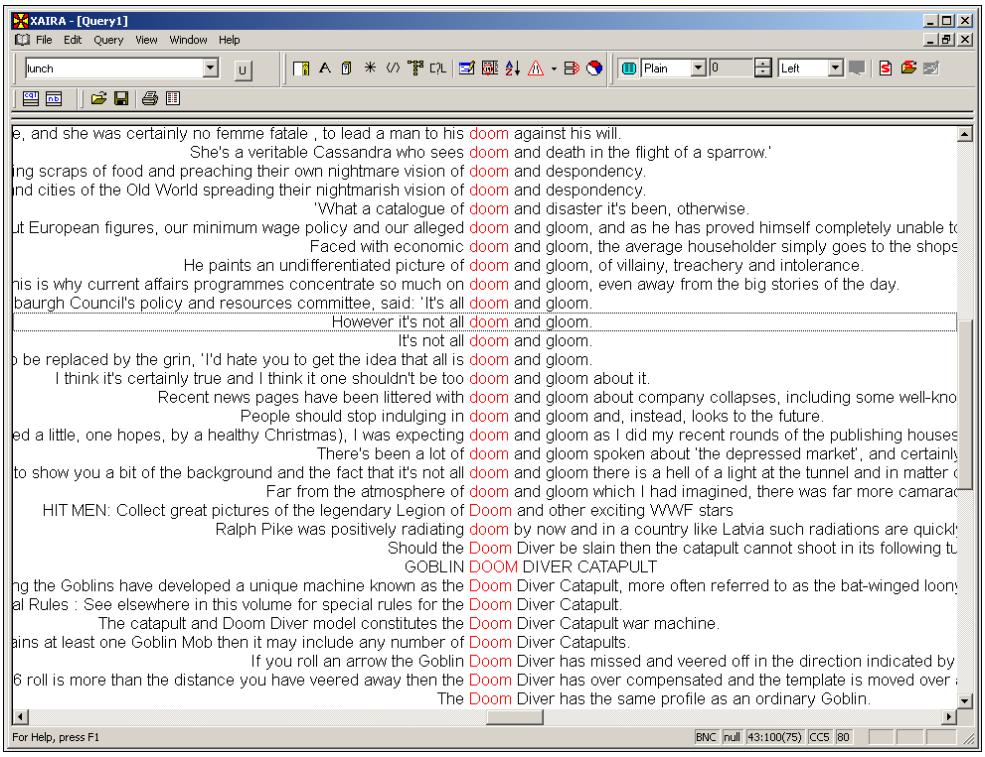

Figure 1: Part of KWIC concordance for ***doom*** in BNC, plain text view sorted by right context

The set of functions which the modules of a XAIRA system provide correspond, naturally, with a

specific research agenda, that of corpus linguistics. Many systems for searching or displaying XML texts tend to do so according to what one might term the book reader paradigm, prioritizing such functions as the ability to display text as a continuous narrative, or to navigate from section to section of the text, or to abstract the narrative structure of a text. By contrast, XAIRA, and systems like it, tend to prioritize a non-narrative view of the text, focussing on its lexis, and on the contexts in which that lexis is presented. Characteristic of this approach are the modes of reading a text typified by the KWIC Concordance (figure 1), which presents the user with a view at the text at once fragmented, in that each occurrence of a word is presented in isolation, and holistic, in that every occurrence of a given word can be compared with every other. More radically still, taking to heart J R Firth's much-quoted maxim that *you shall know a word by the company it keeps*, systems like XAIRA offer the facility to list for a given word those words which appear more frequently in its company than might be expected, according to a variety of statistical predictors (see figure 2).

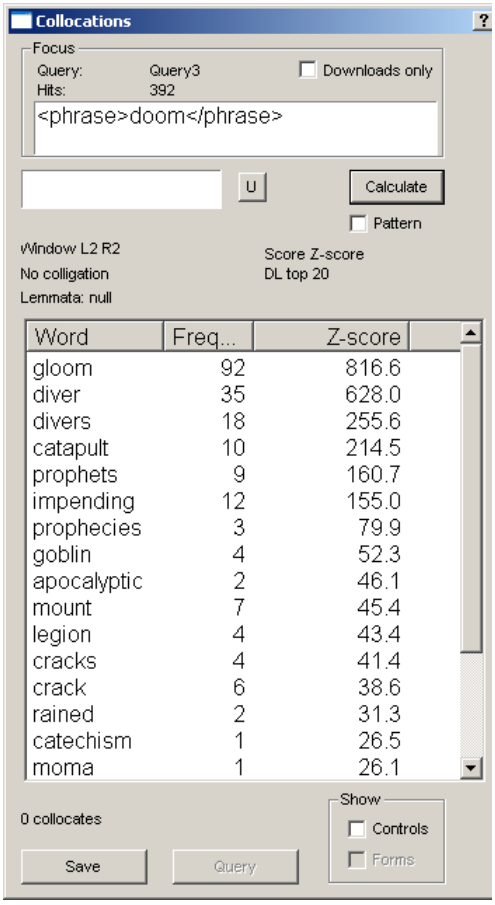

Figure 2: Collocates of ***doom*** in BNC, in window of 2 words L and 2 words R, ranked by Z-score

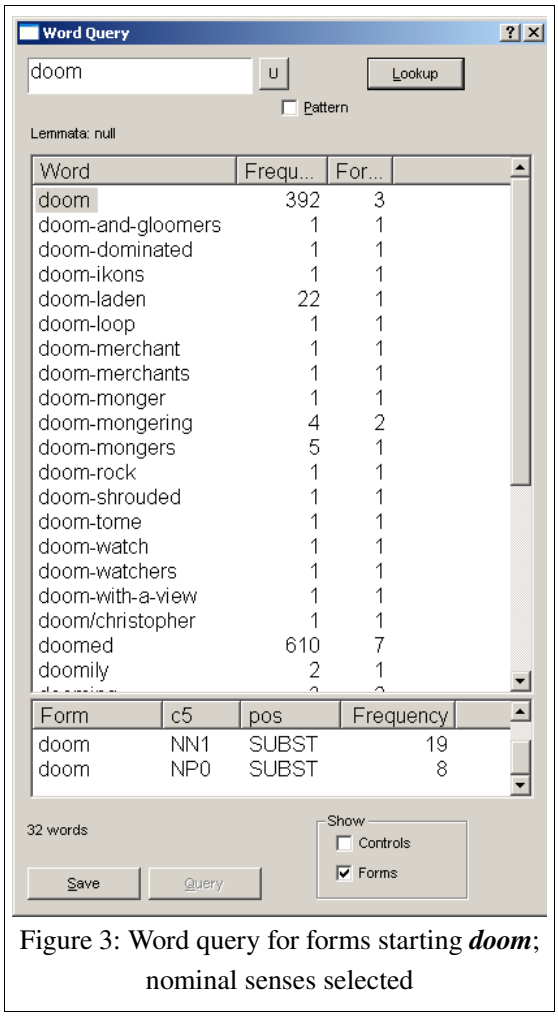

Figure 3: Word query for forms starting ***doom***; nominal senses selected

Where corpora are annotated with additional information, such as part of speech coding, or in which homographs have been disambiguated, even simple word searches can be more sophisticated. In figure 3, for example, we note not only the productiveness of *doom* as a prefix, but also the fact that in a third of its nominal appearances it functions as a proper noun.

Words constitute discourses or texts and it is one of the claims (or at least one of the preoccupations) of corpus linguistics that lexis is one of the ways in which texts can be categorised. A major part of

the discipline is therefore concerned with the partitioning of corpora, and contrastive examination of the lexical properties of the resulting subcorpora. A tool such as XAIRA can help in this process, particularly when the statistical information demonstrating that doom is more characteristic of (say) imaginative and informal writing than it is of academic prose is supported by a simple visualisation such as that of figure 4:

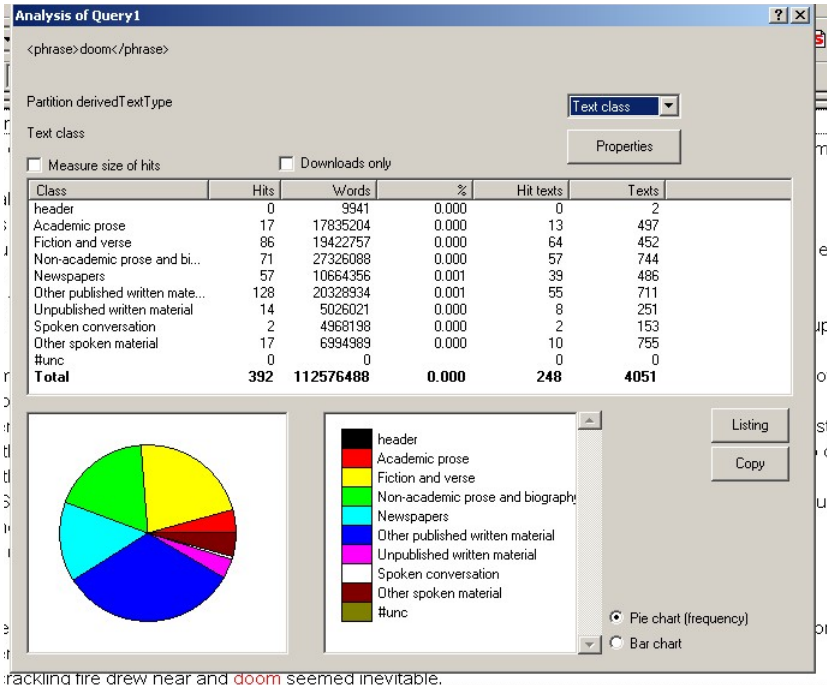

Figure 4: Distribution of ***doom*** across different kinds of text in the BNC

This paper will aim to provide the reader with a reasonably complete overview of how these and other query functions are presented to the user of Xaira, how they be customized for different kinds of corpus, and how their functions may be combined using the different components of the system.

This emphasis on modularity, that is, the separation and specialization of computer processing into discrete components operating to well defined norms and interfaces characterizes today's software development environment, seen most clearly in the development of the Web which both epitomizes and depends upon this trend. But it is also, I shall argue, appropriate to the discipline of corpus linguistics, which, as an inherently empirical and atheoretical pursuit, has always had a common core of functional requirements instantiated to a greater or lesser extent by different software systems. This paper will thus explore these common functional requirements, attempting both to assess how they arise naturally from the research agenda of corpus linguistics, and to describe the extent to which the current version of Xaira supports them.