TGDW16
Software for Searching Large SGML Textbases

Proposal for a research post to be funded by the British Library
Submitted by OUCS and UCREL on behalf of the BNC Consortium

9 August 1993

1 Context

The British National Corpus (BNC) project is currently constructing a 100 million word corpus of modern British English for use in linguistic research. Collaborators in this pre-competitive initiative are Oxford University Press (OUP), Longman Group UK Ltd., W R Chambers, Lancaster University's Unit for Computer Research in the English Language (UCREL), Oxford University Computing Services (OUCS), and the British Library. The project has received substantial funding from the UK Department of Trade and Industry and the Science and Engineering Research Council within their Joint Framework for Information Technology. The current funding period expires at the end of March 1994.

By this date, the BNC project will have available for use within the research community a very large corpus of modern British English, in all its varieties, both spoken and written. The corpus will be marked up and linguistically annotated in a consistent manner throughout, using a simple application of the Standard Generalized Markup Language (ISO 8879: SGML) which is compatible with the emerging recommendations of the international Text Encoding Initiative.

SGML is being very widely applied throughout the information processing industries. Major scientific and reference publishers have adopted it, as have organizations such as ISO itself, HMSO, the EC and the US Department of Defense. Major software vendors (such as WordPerfect Corporation) are developing and marketing SGML-aware software, for the most part targetted at the document production industry, although commercial software for electronic text manipulation is also becoming increasingly available. In the Libraries community, the importance of SGML has long been recognized: some of the earliest UK studies of SGML were funded by the British Library.

Awareness of SGML within the scientific research community is beginning to spread, largely as a result of highly successful applications such as the World Wide Web; the EC Language Research and Engineering programmes (which recommend SGML as an interchange format) have lead to a equal awareness of its importance within the linguistic research community.

The BNC can only benefit from the existence of this wider community of SGML aware users and developers. However, in the short run, the project has a pressing need to obtain or develop in house a simple set of software tools for basic text retrieval and management of a very large SGML document. The current proposal addresses that need.

2 Goals

The chief deliverable of this project will be a package of text-searching software tools, capable of dealing efficiently with large-scale (1Gb+) SGML text bases. The exact functionality of the suite will be defined by the needs of the BNC Consortium, but will include such features as the following:

- Keyword-in-context display
- Retrieval by exact or partial string matching
- Retrieval by word-class tag or syntactic code (where available)
- Retrieval in terms of the SGML document structure

The package should be designed and implemented as far as possible in a machine-independent manner. It is probable that the initial implementation will be for UNIX workstations running under Motif. However, well-documented interfaces to the package will be provided, enabling users to extend its capabilities for other environments or applications.

As far as possible, any specialised software and documentation developed within this project will be distributed together with the Corpus. It is intended that the software should be made as freely available as possible. To that end, a report on the feasibility of achieving the desired functionality by using available public-domain SGML-aware and related software will also be produced.

3 Suggested Functionality

The following set of requirements will be greatly refined and expanded during the initial stages of the project, in close consultation with the BNC Technical Committee. In general the intention is to provide functionality no less than that initially proposed by the IS&RP deliverable as defined in the Consortium Agreement.

- arbitrary user-defined subsetting of corpus in terms of information in the headers or content of text
- KWIC concordance generation
- wordlist and frequency counts
- string-based (exact and fuzzy) retrieval and browsing of lexical and grammatical items
- retrieval in terms of mutual cohesion (”Z-scores”) or other statistically defined criteria
- retrieval browsing and display in terms of SGML structure

Efficient indexing or other techniques will be needed to ensure that performance of the software is acceptable. However, it is not intended that this project should re-invent existing and well-understood software techniques for handling large textual files. Rather, it should seek to apply them within the SGML context.

4 Method

1. A Software Engineer will be appointed as soon as possible to join the existing BNC team at OUCS, but working closely with design and development team at UCREL. It is hoped to have someone in post by October 1st at the latest, and preferably earlier. In view of the time constraints, it is hoped to appoint someone with detailed practical knowledge to develop the system in a short time (six months is proposed) rather than employ a less skilled person for a longer period.

2. A survey will be made of existing public domain or suitably licensable software libraries for relevant tools. These to include among others: SGML parsers, text indexing software, text browsers, text editors. This may be carried out in conjunction with the SGML Project at the University of Exeter.

3. In close consultation with other Consortium members, a modular system architecture for the initial package will be designed. Appropriate pre-existing components will then be selected, either from public domain or appropriately licensed sources. Any necessary additional components will be designed, implemented and tested.

4. In consultation with other partners, and more widely with potential users of the BNC, a set of corpus-specific applications, test queries etc. will be prepared and tested against as much of the Corpus as is available at this stage in the project.

5. Usability and software problems identified by the experience gained in the previous stage will be fixed. The scope and interfaces of the software will be documented. An initial beta test release should be available by February 1994.

6. A report on the results of the survey carried out at (2) will be produced, assessing the viability of the approach taken and identifying any particular problems in the approach taken to searching large text bases.

7. The basic software tools developed by the project will be made available as widely as possible. It is planned to distribute them together with the BNC itself, and also independently of it.