# PCR17
# BNC: Progress Report : 1991, fourth quarter

Lou Burnard

10 January 1992

- *Task group and other related meetings.* OUCS staff attended meetings of Task Group A (21 Oct), C (12 Nov and 10 Dec) and D (5 Sept). There was also some discussion with OUP about revision of project milestones in view of the considerable slippage resulting from delays in agreement on encoding formats and selection criteria.

- *Computer facilities.* No changes in the hardware during this quarter. Direct connexion to the international Internet became a real possibility at the end of this reporting period. We are now considering the security implications.

- *Software.* The public domain SGML parser continues to be of considerable usefulness. We also took delivery of the XTRAN software system produced by Exoterica Corporation in December, made available at a substantial discount to the project under special licensing arrangements. This software provides powerful facilities for converting to and from SGML, together with an excellent SGML parser.

- *Database.* Development of the database has continued following the implementation of a trial system. Current work centres on the representation of text selection and classification criteria in order that reports on the balance of the material in the corpus may be generated.

- *Text Accession.* A sample body of texts totalling a million words in prototype CDIF format had been received from OUP by the beginning of December. A detailed evaluation of this material revealed a variety of discrepancies, documented in TGCW23, which are in the process of being resolved. About half of the texts have now been fully verified.

  Following signature of an agreement between Longman and OUCS, we received several samples from the Longman/Lancaster Corpus, for eventual inclusion in the BNC and release to other participants, when appropriate permissions have been obtained. Some sample spoken texts have also been received. No progress on converting the written materials to CDIF has yet been made as this is not a time-critical task. A specification for software to convert automatically from the Longman spoken text format to CDIF is to be drafted.

- *Text Encoding.* At the start of this reporting period, a preliminary version of the CDIF DTD was tested against a small number of written texts leading to some revision of the DTD. A new DTD, providing all and only the tags agreed to by Task Group C, has now been drafted and is being tested.

  A first attempt was also made at defining a TEI-conformant header structure for the corpus; this however is in need of substantial revision both because of changes within the TEI recommendations and to incorporate extensions for spoken texts.

  Following agreement of the encoding scheme for spoken texts by Task Group C, CDIF was expanded to include TEI-conformant versions of the encoding proposed for spoken texts. A new document providing a definitive version of the whole set of CDIF tags is in active preparation.

- *Text Enrichment.* A set of codes for linguistic annotation was proposed by Lancaster in September, and a preliminary set of equivalent TEI feature-structure declarations was drafted. Following revision of the Lancaster tagset as a result of discussion with other members of the SALT community, the development of a full TEI FSD can now proceed.

- *Documentation.* Aside from minutes and internal notes, OUCS produced working papers on *The BNC in Ireland* (BNCR12), *The BNC in Northern Ireland* (TGAN17), *Text selection and classification criteria* (TGCW20), *Sample CDIF-encoded texts from OUP* (TGCW23), and *The BNC's data catchment area: A proposal* (PCW16).

- *Presentations.* OUCS staff attended the "Using Corpora" conference in Oxford in September and the "Computers and Teaching in the Humanities" conference in Durham in December. During October, a presentation about the BNC was given by LB at "SGML '91" in Providence. In November, GB presented a paper at "Text Retrieval '91" in London.

- *Visits.* In November, GB visited a number of sites in Ireland on behalf of the project, giving lectures and soliciting material, as reported in BNCR12. Partly as a result of this trip, OUCS is promoting a proposal, PCW16, that the BNC should include materials from the whole of the British Isles, not just the UK.