PCA34
BNC: Progress Report : 1992, fourth quarter

Lou Burnard

11 January 1992

- *Computer facilities.* A new disk drive was added to the OUCS system, increasing its total capacity by 2.3 GB. It is anticipated that one further drive of this capacity will need to be added before the end of the project. One of the older drives failed and was replaced, thus demonstrating the value of the maintenance contract and the effectiveness of the back-up scheme.

- *Personnel* No changes.

- *Text Accession.* A total of 28,039,901 words of written text have now been received from OUP, of which 483,139 were returned as insufficiently British, lacking in permissions or inadequately marked-up. So far this quarter, 7,475,521 words have been processed and sent to Lancaster, (nearly twice as many as in the last quarter), bringing the total to 15,860,539. We have still not received any CDIF material from Longmans (either spoken or written). In mid-November, we began producing an automatic status report giving weekly through-put counts in various categories; this is automatically distributed to all bnc-discuss subscribers by e-mail.

- *Text Encoding.*

  With the appearance of CDIF 1.2 at the start of this quarter, no further changes are anticipated in the dtd, other than minor bug fixes, and possible extensions for the "core" corpus (see further below). About 3 person/months this quarter were spent working with or on behalf of other partners to ensure CDIF-compatibility of their texts. A TEI-conformant content model for text headers has been agreed, but has yet to be fully implemented and applied.

- *Text Enrichment.* Word class tags and segmentation strategy for the bulk of the corpus have been agreed and implemented. Work continues on the characteristics and representation of the text enrichment proposed by Lancaster for the "core corpus": an extended set of word-class tags has been agreed, and a set of segment types proposed. Neither has yet been integrated into CDIF.

- *Text Dissemination* Accounts on the BNC system are being provided for participants as required. It had been intended to distribute a "sample tape" at the end of 1992. At the time of writing, this had not been done, because, while there is plenty of written material from OUP which could be redistributed to participants, only small quantities of test data have been received from Longman (and, of this, only the spoken material can be redistributed), and similarly small quantities of word-class tagged material have been received from Lancaster. The views of participants as to the usefulness of the compilation of a sample under these circumstances are being sought.

- *Documentation.* Aside from minutes and internal notes, OUCS produced working papers on *The New BNC Database* (TGCW36) *weeklySummary – summarize corpus throughput* (TGCW40) *minimize – minimize CDIF tagging* (TGCW41)

  Publications included: 'Report on the 13th ICAME conference' *Computers & Texts*, October 1992 *Encoding the British National Corpus* (paper for ICAME proceedings) (BNCX27)