PCA28

BNC: Progress Report : 1992, second quarter

Lou Burnard

19 June 1992

- *Computer facilities.* No significant changes in hardware occured during this quarter. A number of special purpose programs are being developed for automatic conversion into CDIF from OUP and Longman data capture formats. Some delay is caused by the absence of a formal description of OUP's markup format, as required by TGCW27. A special purpose validation program for Longman's internal use has also been developed. Development of some CDIF-specific software proceeds: in particular a utility to summarize tag usage is now available and documented.

- *Personnel* We are currently recruiting an editorial assistant to work on correction and proof reading. Once this person is in post (hopefully early next quarter), throughput should increase significantly.

- *Text Accession.* A total of 12,174,303 words of written text have now been received from OUP. During this quarter, we received about one million words from OUP, and just under a million words of written materials from Longman. So far this quarter, just over four million words have been processed and sent to Lancaster.

- *Text Encoding.*

  Work is in progress on the CDIF header, and should be complete by the time of the PC meeting. As more varieties of text are processed we are also identifying a number of minor improvements which could be made to the current version of CDIF. These will be proposed to Task Group early in the next quarter for approval, and implemented thereafter assuming that other work permits.

- *Text Enrichment.* Discussion with Lancaster has indicated some problems in the accuracy of texts sent to them for analysis: steps to address these problems are being taken. We have not yet received any enriched texts from Lancaster and have therefore been unable to assess the effectiveness of the CDIF representation for word class tagging.

- *Text Dissemination* Work has commenced on providing means of making texts available to project participants as rapidly as possible.

- *Documentation.* Aside from minutes and internal notes, OUCS produced working papers on *A program for summarizing* CDIF *tag usage* (TGCW31), *Using* REFORMAT *and* CHECK (TGCW32), OUCS ALLC *foils* (BNCP20), and OUCS ICAME *foils* (BNCP24).

- *Presentations.* In April OUCS staff attended the ACH/ALLC conference. A special session in which all BNC participants were involved was well attended and well-received: see document BNCP20. In June, Burnage and Dunlop attended the ICAME conference in Nijmegen and gave a joint presentation on the project: see document BNCP24.