PCA31
BNC: Progress Report : 1992, third quarter

Lou Burnard

1 October 1992

- *Computer facilities.* There were no changes in hardware during this quarter, but we are planning to increase the amount of disk storage at our disposal by purchasing an additional drive. Software development has focussed on the text management database which will be used to monitor work flow and provide management information; in addition an interactive routine for dealing with hyphenation in incoming texts has been developed.
- *Personnel* Glynis Baguley joined OUCS as full time editorial assistant at the start of August. With her help, we hope to raise the throughput of texts significantly during the next quarter.
- *Text Accession.* A total of 23,324,317 words of written text have now been received from OUP. During this quarter, we received about five and a half million words from OUP (a further 2 million words arrived at the time of writing); no additional materials were received from Longman. So far this quarter, 4,769,589 words have been processed and sent to Lancaster, bringing the total to 8,385,018.
- *Text Encoding.*

  Great progress has been made on defining the content for the CDIF header, following publication of the TEI's Recommendations for the Header in August. Some additions needed in the CDIF specification were submitted to Task Group C for approval at the end of September. Work is in hand to convert the whole of the backlog of texts to this new specification. Several technical problems in getting texts (both written and spoken) from Longmans in a format that can be converted to CDIF have been resolved, following discussion and a visit to Longmans by OUCS staff.
- *Text Enrichment.* We have now received detailed specifications for the word class tagging to be carried out at Lancaster, and also preliminary information about the skeletal parsing. These will be incorporated into the CDIF documentation at the next revision.
- *Text Dissemination* Other project participants with access to the Internet will be provided with an account on the OUCS machine on request. More elaborate provision for access to the database has been deferred until the likely recommendations of the Exploitation Committee are clearer.
- *Documentation.* Aside from minutes and internal notes, OUCS produced working papers on *Is the conversion of Longman/Lancaster texts to CDIF possible?* (TGCW26); *Corpus Document Interchange Format v. 1.2* (TGCW30); *Relationship between the TEI.2 header and the BNC corpus and text headers* (TGCW34); and *Corpus text processing: directory structure and file names* (TGCW35).
- *Presentations.* LB spoke about the BNC as an example of the TEI recommendations at a pre-COLING Workshop, and also consulted with the French part of the Network of European Research Corpora. The project received two sets of visitors from the far East: Profs Kitamura and Ken Horii from Kansei University, Japan; and Profs. Liu Lianyuan and Fu Yonghe from the Department of Language Planning and Processing of the People's Republic of China.