PCR18
BNC: Progress Report : 1992, first quarter

Lou Burnard

25 March 1992

- *Computer facilities.* No significant changes in hardware or software occured during this quarter.
- *Text Accession.* During this quarter, nearly 9 million words were received from OUP. Following agreement and testing of the acceptance procedures described in TGCW27, the process of processing these texts has begun. At the time of writing 2 million words have been through both syntactic and semantic checks and are now being forwarded to Lancaster for word class tagging. Targets for future run-rates have been agreed with the project manager.

  No additional material, either written or spoken, has been received so far from other participants in the project. No further progress has therefore been made on testing ease of conversion to CDIF from other formats: the specification for software to convert automatically from the Longman spoken text format to CDIF is still to be drafted.
- *Text Encoding.*

  Definition of those aspects of CDIF relating to the body of written and spoken CDIF (including the tags used for marking parts of speech) is complete. The first draft of a full specification for CDIF (TGCW30) was completed this quarter, for approval by Task Group C and the Project Committee. This currently lacks only information relating to the definition of text and corpus headers which should be completed during the next quarter, now that further input on the topic has been received from the TEI.

  A variety of software procedures have been developed and tested to simplify the running of the syntactic and semantic checks so far defined.

- *Text Enrichment.* The revised set of word class codes received from Lancaster has now been mapped onto equivalent TEI feature-structure declarations. Further testing has been delayed, pending the receipt of some enriched texts from Lancaster and the results of discussions of the codes with other members of the SALT community.

- *Documentation.* Aside from minutes and internal notes, OUCS produced working papers on *Markup for non-ISO 646 invariant part characters* (TGCW25), *Corpus Acceptance Procedures* (TGCW27), *Corpus Document Interchange Format* (TGCW30), and *Feature structure declarations for the BNC Wordclasses* (TGDW09).

- *Presentations.* During January, OUCS staff attended a workshop on lexical resources in New Mexico, at which LB made a brief presentation concerning the BNC jointly with Steve Crowdie); LB and several other BNC participants also attended the long-awaited Corpus Workshop in Pisa.

- *Data Protection Act* The British National Corpus Project will, when complete, hold personal information which falls within the scope of the Data Protection Act. (Author names; user names and possibly contact information...) Following advice from Oxford University's Data Protection Officer, the separate registration of the Project has been put in hand.