TCM02
BNC Technical Committee
Minutes for meeting of 25 February 1993

Lou Burnard

1 March 1993

Present:

RA	Bob Allen	Chambers
MB	Michael Bryant	Lancaster
LB	Lou Burnard	OUCS
SC	Steve Crowdy	Longman
DD	Dominic Dunlop	OUCS
LE	Liz Eyes	Lancaster
FK	Frank Keenan	OUP
RW	Ray Woodall	OUP

1 Administrative

There were no apologies for absence.

MB took the chair; LB volunteered to take the minutes. It was agreed to re-order the Agenda, taking items in the order 9, 10, 12, 7, 5, 6 and 8. Item 11 (Regional sampling points in spoken texts) was held over. SC placed an item on Publicity under AOB.

2 Minutes of previous meeting

Accepted. DD had circulated a list showing status of (most) outstanding OUCS action items. Status of all items noted as open or ongoing in TCM01 is as follows:

- RW to investigate cost and feasibility of further British Library funded research into classification of BNC texts. *Closed* (see 3).
- SC to obtain review of dialectal forms used in transcribing spoken materials from John Wells. Review is in progress; SC to circulate a copy of the control list for dialectal forms as soon review complete. Concern was expressed that important socio-linguistic information about respondents would be lost if not included with the "header" information supplied by Longman. Concern that this action item had been open for over six months was also expressed. Open.

SC

- LB to make proposal to Project Committee on storage of original texts. This had been discussed at the last PC meeting, with general agreement that access to the original printed texts was desirable, but no clear indication of how it might be provided in the long term. OUCS would continue to store originals. *Closed*.

- DD,RW to discuss use of <HEAD> and <CAPTION>: ongoing. Document TGCW46 to be produced. Open.

DD, RW

- LB to revise TGCW27 to reflect changes in TGCW30. Open.

LB

- DD,RW to discuss data transfer issues. No progress since last meeting. Open.

DD, RW

- LB to revise TGCW30 to include contents of TGCW34. Some progress was made after the meeting. Open.

LB, DD

- SC, RW to quantify likely shortfall in corpus accessions caused by permissions difficulties. RW noted that the hoped-for change of heart at Reed International had not occurred, and that the Routledge Group and Fraser Dunlop (a major IPR agency) were also now proving unwilling to co-operate. SC noted that blanket refusals from such agencies might have considerable impact on the proportion of material from the Longman Lancaster Corpus which could be included in the BNC. Ongoing.

SC, RW

- DD to generate sample tape for participants unable to gain access to OUCS machines. DD asked what kind of medium was required, expressing strong preference for cartridge. SC said that standard half-inch reel tape for VMS would be best. LB said this was possible. Noted that this was a sample for demonstration purposes only; as previously stated at Project Committee meeting, it would not be available till after Easter. *Closed.*

- RW to investigate further sources of machine-readable newspapers. Recent successes included the South Wales Evening Post (RW suggested that this might have been triggered by a recent Welsh sporting triumph); the Scotsman and the Shropshire Star as well as other publications such as Unix News and White Dwarf. Ongoing.

RW

- LE LB to arrange BNC presence at SERC/JFIT poster session. Lancaster would be sending Nicholas Smith to the meeting to be held 23-4 March at Keele. OUCS might send someone in addition. *Closed.*

- LE to draft proposal on error handling at Lancaster. A report (TGCW44) was tabled at the meeting. See 10 . *Closed.*

- DD, RW to discuss linebreaks in headings. Topic is addressed in TGCW45. *Closed.*

- LB to update TGCW30 to reflect current C5 tagset. Agreed that revised document would not be circulated until other pending changes complete. *Closed.*

- LE to propose further portmanteau tags if shown to be necessary by frequency analysis of CLAWS output. Experience over the last four months indicated that further tags would be unlikely, but Lancaster wished to reserve that possibility. *Closed*.
- DD to add C6 tags to CDIF dtd. This would be subsumed in the ongoing work on updating TGCW30. *Closed*
- LE to circulate revised TGDW11, possibly with shorter tagnames and changes to treatment of punctuation. *Closed*.
- LE to document labels used for constituent segments in output from skeletal parsing and to provide samples of such material to OUCS. The content of TGDW12 had now been transferred to document TGDW14 and included an example. *Closed*.
- LB,DD to extend CDIF to provide support for skeletal parsing. See 8 .
- RA to provide detailed proposals for collection of miscellaneous unpublished materials. These had been produced but not yet formally presented. See 7 .
- MB, DD to discuss ways of identifying postedited segments. Ongoing. See 4 .
- DD, RW to discuss keyboarders' instructions for unpublished material. Document TGCW46 will address this. Open.

DD, RW

- DD to propose algorithm for determining Britishness of a text. Document TGAW22 describes the proposal. *Closed*.
- RW to establish OUP's position with respect to contribution of software to the IS&RP. No progress. Open

RW

- MB to update PCW32 in light of discussion of IS&RP. *Closed*. (see 9)
- DD to provide MB with list of X-Window application builder products. *Closed*.
- MB, DD to discuss retrieval aspects of IS&RP. *Closed*. See 9 .
- MB to canvas potential users for wish list of features in IS&RP. *Closed*.
- FK, RW to provide wish lists of IS&RP features. Open.

FK, RW

- SC to report on ways of achieving the target size for the spoken components of the BNC. *Closed*. Document is PCW40.

3 Interim Progress Reports

RW reported for OUP that so far six of the 11 million words targetted for this quarter had been scanned or keyed. Delays in obtaining permissions had lead to some problems. Machine-readable data was coming in, but less than had been promised and (as anticipated) with conversion problems, most of which were being satisfactorily addressed. It was important to expand input in machine readable form (e.g. to include books) since the scanners were now used to full capacity. R. Sweeney would be making a further set

of recommendations for titles to be included. A visit from the University of Barcelona was expected.

SC reported for Longmans that the throughput of the speech transcribers remained satisfactory and that the problem in treating overlapping speech had been resolved. Inconsistencies in the use of certain tags (notably <NV>) remained problematic; SC had discussed these with OUCS staff. Two new transcribers had been recruited from COBUILD with particular responsibility for Northern region data. Collection of the context governed data continued, most recently featuring six hours from the mayor of Cambridge. Della's recent radio appearance had generated much interest.

DD reported for OUCS that a peak rate of 1.2 million words a week had been achieved but only on "good stuff", of which there was none left. Other commitments also made it difficult to keep up this rate. A lot of magazine material would be returned to OUP shortly, largely because of problems related to the encoding of headers and divisions, and spelling errors. The 16 million target for the quarter was unlikely to be achieved unless the quality of material received was significantly improved. DD had submitted a paper for presentation at ICAME in May. Gavin Burnage and Roger Garside would be making a joint presentation at ACH-ALLC in June.

RA had nothing to report for Chambers, as yet.

MB reported for Lancaster that about 3 of the 25 millions targetted for this quarter had been processed; the remainder should be ready for delivery within three weeks. Much time had been spent adjusting their operational procedures with a view to accelerating throughput. The proportion of texts retained for post-editing would be reduced to 1 in 5. A delivery method enabling Lancaster to "claw back" material where necessary had been worked out. So far they had only worked on "well-behaved" texts using the C5 tagset. Upublished material might present new problems, although they had already dealt with some, e.g. DHSS leaflets. Spoken materials were also potentially problematic.

4 CDIF Specification (TGCW30)

For rather obscure reasons, the original CDIF had included floating <LABEL> elements, which had been found to cause serious problems at segmentation time. Standard practice was now to use the <OMIT> tag in such situations. The CDIF specification would be updated accordingly.

Lancs requested a means of indicating within CDIF that a given <S> or block of segments had been postedited. After discussion it was agreed to introduce an attribute indicating the status of a segment for this purpose.

LB *to provide support for segment status in* CDIF

5 Data transfer format for unpublished materials

This would be resolved between OUP and OUCS since Chambers had delegated the task of data capture to OUP. A format would be decided on as soon as possible.

6 Britishness (TGAW22)

Discussion centred on whether or not the presence of US spelling should be given a higher weight than information about the authors. The former required access to the source, whereas the latter required an unpredictable amount of background research. There was general agreement that the cost of the book would generally be less than the cost of the research. It was also noted that at present more texts than anticipated were

failing the Britishness Test. RA and others noted that "US spelling" was not an entirely water-tight notion in any case; however some defensible and deterministic procedure was clearly necessary. It was finally agreed that OUP would apply the test as specified, and that OUCS would continue to check its application.

7 Miscellaneous Unpublished Material

RA reported that a revised proposal for collection of up to 4 million words, in part modelled on the schema adopted by Longmans for the spoken component, was now with OUP for consideration. The scope had been broadened to include low-circulation printed material such as church magazines. OUP already had about one million words of such material, the fate of which was unclear. RA was requested to circulate this report (allocated BNC document number TGAW23) and a further draft which describes the various categories of source material to be collected (TGAW24).

It was noted that Chambers would focus on data collection, with data capture being performed by OUP. With reference to the proposed "demographic" method, SC urged the importance of carrying out a pilot, given the difficulty of predicting likely response rates, and the consequent possible imbalance. RA and RW pointed out that there was insufficient time for this. The selection procedure used for the spoken corpus might be re-usable, but expert opinion (i.e. the BMRB) should be obtained if time and cost permitted. Chambers were still waiting for an official response from OUP; no further planning could be carried out in the absence of a budget or a timescale. RW said that unless further funding was made available, e.g. from the British Academy, the "demographic" part of the Chambers proposal could not possibly be funded. It was agreed that OUP should however address the two parts of the proposal independently.

RW *to provide a formal response to the Chambers proposals for gathering unpublished material*

RA *to circulate drafts of TGAW23 and TGAW24 to Technical Committee*

8 Core Enrichment Proposals (TGDW14)

This paper presents a reprioritisation of Lancaster's tasks in view of the diminished resources now available. First priority had to be the 100 million words tagged in C5, followed by the 2 million 'core' corpus tagged using C6. The proportion of this which would be further enriched was not clear, but was likely to be small. Only skeleton parsing would be provided; prosodic annotation was no longer envisioned. To accomplish these targets, the rate of error checking had already been reduced to 1 block in 5. Further funding if provided would enable this to be raised to 1 in 3 for C5 material, and 1 in 1 for the core corpus.

Skeletal parsing would be carried out only on segments sampled from the core corpus according to a procedure to be defined. Contiguous blocks of about 100 sentences would be chosen from a variety of texts, up to a target of 50,000 words.

The C5 and C6 tagsets are independent, and could therefore both be represented in the same text without ambiguity.

LB said that OUCS would propose a CDIF-conformant representation for skeletally parsed material and would liaise with LE about the best method of achieving it. It was generally felt that this should be independent of the C_ file format.

LB noted that an independent feature system declaration had yet to be defined for the C5, C6 and skeletal parse tagsets.

LB,LE *to discuss* CDIF *representation for skeletally parsed materials.*

LB *to send feature system definition to LE for review*

9 Information Search and Retrieval Package (TGDW15, PCW32)

MB explained that PCW32 had been expanded to clarify the desired functionality of the ISRP as requested in previous minutes, while TGDW15 presented what was pragmatically possible in light of the dwindling resources available. The meeting sympathized with Lancaster's difficultiess in this respect.

LB suggested that searching should be able to address the SGML document structure (ESIS) directly. DD noted that indexing code was both hard to develop and easy to acquire. MB agreed to investigate alternative existing toolkits, where resources permitted; uncertainty as to the form in which end-users of the corpus would acquire it, or how, further complicated the issue.

Progress in development of the ISRP would be monitored internally at Lancaster, with internal reports and decisions. It was intended for demonstration purposes only.

10 Error processing at Lancaster (TGCW44)

There was some inconclusive discussion of the policy issues raised by this paper, chiefly centred on how Lancaster should treat apparent typographic or other errors found in text after parsing. OUCS policy was to mark faulty material correctly transcribed with the <SIC> tag; this was however only possible by reference to the original source and was a very costly option. Lancaster already flag places where human intervention had been involved in the assignment of a word class code and could therefore indicate at least segments where a change has been made to the text during post-editing in order to get a better parse, perhaps by an appropriate value for the proposed postedit status attribute. The CLAWS "decision code" would not however change for all kinds of human intervention; changes in segmentation for example would remain unmarked. Changes in the position of punctuation relative to tags could reasonably be made silently since these were to some extent arbitrary. No conclusion was reached.

all *review TGCW44 and suggest solutions to problems it raises.*

11 Any Other Business

It was agreed that copies of any publicity releases, interviews, etc. related to the project should be distributed internally. An appropriate mechanism would be via the project management structure; copies of all such material should therefore be sent to RW in future, for distribution to other partners.

The meeting closed at 1530. No date was fixed for the next meeting.