TCM03
BNC Technical Committee
Minutes for meeting of 27 May 1993

Gavin Burnage
Dominic Dunlop

11 June 1993

Present:

RA	Rob Allen	Chambers
MB	Michael Bryant	Lancaster
GB	Gavin Burnage	OUCS
LB	Lou Burnard	OUCS
SC	Steve Crowdy	Longman
DD	Dominic Dunlop	OUCS
SMB	Simon Murison-Bowie	OUP
RW	Ray Woodall	OUP

1 Opening of meeting

The meeting commenced slightly after 14:00. DD took the chair, and asked GB to take the minutes.

DD took the opportunity presented by the meeting to distribute new or revised documents TGCW35, TGCW40 and TGCW48–50 for information. The documents were not discussed at the meeting.

It was agreed to re-order the Agenda and deal with item 4 (IS&RP) before all other items as LB and SMB needed to leave at 3pm.

2 Review of actions from last meeting

The minutes of TCM02 were accepted. The status of items shown in those minutes as open or ongoing is as follows:

- *SC to obtain review of dialectal forms used in transcribing spoken materials from John Wells.*

  Ongoing. A preliminary list of these forms, and of the forms used in the transcription of vocalized pauses and the like, has been passed by SC to GB, who has in turn forwarded it to Lancaster. The document number TGCW53 was allocated for this list. However, a review has yet to be received from John Wells. **SC**

- *DD, RW to discuss use of* `<head>` *and* `<caption>`.

  Closed. A new version of OUP's *Corpus markup — codes for keyboarders* has been produced. This document is numbered TGCW52.

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

  Open. **LB**

- *DD,RW to discuss data transfer issues*

  Ongoing. Held over pending OUP's imminent appearance on the Internet. **DD, RW**

- *SC, RW to quantify likely shortfall in corpus accessions caused by permissions difficulties.*

  Ongoing. RW noted that, while both Routledge and Fraser Dunlop had now granted permissions, the Reed group was still refusing despite approaches from within OUP at a high level. He thought that, despite Reed's refusal, OUP would be able to make up its quota. SC reported that the process of obtaining BNC permissions for works already in Longman's corpus was time-consuming, but moving ahead. The impact on Longman's contribution to the BNC of any refusals has yet to be evaluated. In the event that twelve million words could not be made available from material in the Longman corpus satisfying the current BNC sampling strategy, Longman might request the project committee for a relaxation of the publication date requirement. DD expressed reservations about this. Alternatively, Longman might seek permissions on, and capture, material not already in its corpus. (See also §3.2.) **SC, RW**

- *RW to investigate further sources of machine-readable newspapers.*

  Closed. (See §3.5.)

- *DD, RW to discuss keyboarders' instructions for unpublished material. Document TGCW46 will address this.*

  Closed. The group felt that document TGCW53 (see above) was sufficient to close this item. TGCW46 will not be produced.

- *RW to establish* OUP*'s position with respect to contribution of software to the* IS&RP.

  Closed. The proposed new arrangements for IS&RP (see §4) make it appropriate to close this item for the present. It may be reopened when the new arrangements are in place and possible sources of software are investigated.

- *FK, RW to provide wish lists of* IS&RP *features.*

  Closed for the same reason as previous item.

- *LB to provide support for segment status in* CDIF.

  Closed. DD has updated the working version of CDIF to provide a segment status attribute.

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

  Closed. A response has been provided and agreement reached.

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

  Closed. TGAW23 has been renumbered at PCW38.

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

  Ongoing. **LB, LE**

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

  Open pending conclusion of previous item. **LB**

- *All to review TGCW44 and suggest solutions to problems it raises.*

  Open. **All**

3 Short status reports

3.1 Chambers

RA reported that, in order to obtain material for the unpublished part of the BNC, around a hundred potential source organizations, both inside and outside Chambers, had been contacted. Response has initially been slow, but the Scottish Publishers' association had contributed 63,000 words on diskette, and some hand-written material had been obtained. A press release detailing the project and soliciting material had produced little response. Chambers intended to recruit a person to work full-time on contacting institutions, and on following up those contacts. The matter of obtaining BNC permissions will also be addressed.

Anonymity is an issue with unpublished material. Henry Thompson is involved in the production of software which automates the process of removing identifying information. RA will discuss this with him at the forthcoming Advisory Committee meeting in Edinburgh. **RA**

3.2 Longman

SC reported that the transcription rate for spoken material was satisfactory, and that much new context-governed material had been collected. Between four and five million words had been transcribed to date. Some problems were being experienced as a result of applying additional transcription and header checks agreed with OUCS, but these applied mainly to material "in the pipeline": new material should be less problematic.

A new person has been taken on to attend to the mark-up of existing written material, again in a manner agreed with OUCS. SC expected that, owing to difficulties in obtaining permissions and because BNC selection criteria ruled out some of the material in the Longman corpus, a maximum of eight to nine million words would be available. (See also §2.)

Longman was currently re-budgeting its BNC-related activities.

3.3 Lancaster

MB reported that eighteen million words had been tagged by CLAWS, with a further six million awaiting processing pending resolution of a software problem which was expected to be fixed very shortly. Processing of core corpus texts has also begun.

The processing of spoken material required some re-jigging of CLAWS. This done, initial material would be subjected to 100% post-editing prior to delivery to OUCS. MB expected that deliveries would commence before the end of the quarter.

MB pointed out that Lancaster would not meet its milestones for the quarter unless sufficient material was delivered by OUCS to Lancaster.

3.4 OUCS

DD reported that OUCS throughput was somewhat constrained by deliveries from OUP: "easy" material such as books and newspapers was quickly processed, but, when it ran out, the more difficult material (magazines etc.) took longer to process, and resulted in a higher level of "bounces". The quality of difficult material did appear to be improving, however.

GB reported that a steady stream of spoken material was now being processed.

DD noted that headers had yet to be constructed for corpus texts.

3.5 OUP

RW reported that it was taking much effort to achieve the quarter's target of eleven million words, particularly in view of continuing problems in obtaining permissions, and of an office move which had recently taken place. Ginny Frewer, currently responsible for obtaining permissions, is transferring to another position within OUP. Her post will be re-advertised as a matter of urgency.

The quest for machine-readable data has been successful, yielding five million words, of which three million have been processed to date. Partly as a result of this, the project's requirements for newspaper material have already been met. Meanwhile, scanning is a bottleneck: it is likely that a further Macintosh-based scanning system will have to be purchased and operating staff recruited both to run it, and to replace casual staff who have recently left the project.

4 Proposal for Production of IS&RP

This item was taken out of order — see §1.

LB spoke on TGDW16, slightly amended following feedback from project participants. RW said that point 1 under GOALS was in fact a part of the procedure involved in getting the IS&RP, not an end in itself. The meeting agreed to remove this goal from the proposal.

LB noted that proceeding with the proposal required a leap of faith, since if basic public-domain software were not available, the proposed project would fail. He also noted that the results would come in the form of a basic UNIX tool, and that follow-up work outside the scope of the proposal would be required to make best use of the tool on other platforms. The tool would offer a graphical user interface to basic functionality, and, while it would provide scope for extension to address new uses, such extension was again outside the scope of the proposal.

SC stated that, while Longman had, with a Danish company, jointly developed an SGML-aware system for lexicographers, it would not be licenced to competitive dictionary publishers. This, and its current requirement for a networked OS/2 environment, probably ruled it out as a component for IS&RP.

SMB commented that obtaining additional funding for IS&RP— a part of the original consortium agreement — might result in problems with the DTI. However, he had heard nothing against the proposal as yet. Following a suggestion from LB it was agreed that Lancaster, funded by SERC under the agreement to provide IS&RP, should discuss the issue with SERC. There is no reason to believe that SERC would object to additional funding being obtained from an outside source, but, noted RW, there might still be a problem in providing the accounting information required by the DTI. **Lancaster**

The group considered a letter from Della Summers to RW in response to the draft IS&RP proposal, and decided that all the points it raised had been addressed. There was some discussion of licencing arrangements for the software. LB stated that, if for some reason it was not appropriate to put the material into the public domain, it should be made as freely available as possible. It could, for example, automatically be licenced to all those licenced to use corpus materials.

It was agreed that LB should clean up the the proposal by adding the necessary descriptive material about the BNC, then submit it to the British Library as a matter of urgency, copying the final document to the Project Committee. **LB**

5 Problems in filling social & community balance criteria quota

RW distributed a chart showing the amount of text allocated to date to each of the BNC topic categories relative to the ultimate target for the category. Most categories were progressing satisfactorily, but there was too little in *social and community* and too much in *applied science* — the latter being almost full. After some discussion, it emerged that Longman experience, from which the category names and volume targets were derived, had assigned texts in *home economics, family living, medical sciences* and *psychology* to *social and community*, whereas OUP had been assigning such texts to *applied science*. This probably accounted in full for the discrepancies seen by OUP.

It was agreed that RW would write a proposal either to reallocate affected texts to the alternate category, or to change the targets for the affected categories so as to reflect their make-up. In the even that texts are reallocated, OUP must send revised database records to OUCS. **RW**

6 Core corpus and tag set issues

MB proposed two small changes to the C6 mark-up. Firstly, the `MC-MC` tag should be renamed `MCMC`; and, secondly, a new tag, `FU` should be added, corresponding to the `UNC` tag in the C5 tagset. Both changes were agreed. MB will reissue TGDW11 to reflect the changes. **MB**

SC asked whether further tags would be required for spoken texts, for example to handle truncation and voiced pauses. MB replied that CLAWS' existing idiom-list mechanism could be extended to deal with such situations, and than no further tags should be required.

7 Issues arising from distribution of sample tape

DD had distributed copies of a sample tape carrying approximately twelve million words to the commercial partners. No issues arising were reported.

8 Computer requirements for Chambers

Chambers is equipped with networked personal computers; it has no UNIX systems. RW sought advice from the group on the tools available to manipulate corpus data. SC noted that Longman also used networked PCs, but these were running OS/2, not MS-DOS. The dictionary system used by Longman was commercially available, but the corpus system was not. (See also §4.)

After some discussion, it was agreed that RW and Chambers' computer officer would arrange visits to Lancaster and OUCS for further discussions. **Chambers, Longman,** OUCS

9 Longman word counting methodology

The matter of the manner in which words are counted in spoken texts had been referred to the technical Committee by the Project Committee (see PCM43). In subsequent discussion between Longman and OUCS, it had been agreed that the following method was acceptable for files passed by Longman to OUCS.

1. Remove header from file.

2. Obtain word and line counts using WordPerfect.

3. Subtract line count from word count to give BNC word count. (This compensates for the counting of <u> tags in step 2, as there is one <u> tag per line.)

The resulting word count is a little higher than that given by OUCS' method, which ignores all tagging delimited by angle-brackets, not just u tags. However, a test at OUCS shows the difference to be only 2–3%, and it was agreed that this was not worth worrying about. The group concurred.

10 Data protection issues

DD reported that the Data Protection Registrar had declined to register the BNC project under the Data Protection Act, as the term "project" was considered to refer to too fleeting an entity for registration. DD will contact the registrar in an attempt to resolve the situation. **DD**

SC asked whether corpus users would have to register as data users under the data protection act. DD considered that those at academic institutions would almost certainly be covered by their institution's registration. The situation of individuals was less clear.

11 Any other business

SC noted that some of the material in the Longman corpus had permissions for a particular region, rather than the whole world, and asked whether this was acceptable. RW commented that, while OUP had endeavoured to obtain worldwide permissions for all its texts, it had had to be content with less in some cases. Provided that European permissions could be obtained, OUP considered a text suitable for inclusion in the BNC. It was agreed that Longman should follow this guideline.

DD noted that OUCS' database was able to record the region or regions in which permissions had been obtained, the rights holders in each region, and any special clauses which should be quoted in connection with a particular grant of permissions. However, no information of this type had yet been entered: arrangements should be made with data capture agencies for its provision in electronic form.

12 Review of agreed actions

Actions are as indicated by notes in the margin of these minutes.

13 Date of next meeting

No date was fixed, but it was agreed that the meeting should be hosted by Chambers in Edinburgh.

14 Close

The meeting closed at 16:40.