TGCM29
BNC Task Group C
Minutes for meeting of 19th February, 1992

Dominic Dunlop

26th February, 1992

Present:

GB	Gavin Burnage	OUCS
LB	Lou Burnard	OUCS
JC	Jeremy Clear	OUP
SC	Steve Crowdy	Longman
DD	Dominic Dunlop	OUCS

1 Opening of meeting

The meeting convened at 14:30. DD agreed to write the minutes; LB chaired the meeting.

It was agreed to discuss arrangements for the ALLC panel session, and the production of a BNC information mailing package under §**??**, Any Other Business.

Document TGCW28, the CDIF DTD, was distributed.

1.1 Minutes of last meeting

> Owing to an omission from the agenda, for which I apologise, the minutes of the meeting of 10th December 1991, TGCM22, were *not* reviewed. What follows is my understanding of the status of the action items from those minutes. If this information is incomplete or incorrect, please let me know, and I shall reissue these minutes.

1.1.1 Action item status

MB *Provide final definition of part-of-speech tags.*

Done. See TGDW08, version of 5th February, 1992.

GB *Draft formal proposal to Project Committee on broadening of scope of* BNC *to include material from the whole of the British Isles.*

Done. See PCW16.

GB, DD *Continue study of sample texts from* OUP *and Longman; report on findings prior to Project Committee meeting of 14th January.*

Ongoing. TGCW23 reflects findings on the OUP texts. No report has yet been circulated on Longman texts — see §**??** below.

LB, DD *Deliver specification of* CDIF *omiting text and corpus header contents.*

Done. See TGCW28 and §**??** below.

JC *Pass copy of revised permissions document to* OUCS.

Done.

JC, SC *Compare contents of* OPC *and* LLC *with a view to eliminating duplicates from material to be proposed for inclusion in* BNC*; make joint approach to copyright holders seeking* BNC *permissions on selected texts.*

Open. **JC, SC**

SC *Mail updated version of TGAW14,* Spoken Corpus Design Specification*, to members of task group A.*

Done.

SC *Approach Clive Upton and John Wells for authoritative opinions on the permitting and representation of "distinct dialectal forms" in transcriptions of spoken material; draw up and circulate an initial control list of such forms for review and comment.*

Open. **SC**

SC *Update §12 of TGCW21 to reflect decision on representation of acronyms recorded in TGCW19.*

Done.

SC *Provide detail of division of country into three regions for purpose of determining balance in spoken corpus.*

Open. **SC**

SC *Provide costings for extending spoken corpus data collection to whole of Ireland.*

Closed. This was not done, but is now moot, as the proposal of PCW16 was rejected by the Project Committee.

DD *Issue revision of TCGM19, the minutes of the meeting of 12th November, reflecting agreed amendment.*

Done.

DD *Issue revision of TCGW18,* Corpus Header*, reflecting new input.*

Open. Work can now begin on closing this item, held open from the 12th November meeting, as the specification of the TEI header is nearing completion. **DD**

2 Proposed OUCS acceptance procedures

LB presented TGCW27, asking for responses from the group on each section in turn. In general, the group accepted the proposals, but requested the amendments listed in the following paragraphs. A new version of TGCW27 is to be produced and circulated to the group. Subject to resolution by mail, phone, or facsimile of any remaining issues, the resulting document will be forwarded to the Project Committee with a proposal from task group C that it be adopted. **LB**

The following minutes present the issues which arose in discussion. Where the group concurred with a section and made no substantive comment, this is not minuted.

§**3 — Delivery formats.** JC asked whether OUCS expected OUP to provide a definition of its delivery format, which is close to CDIF. LB replied that it did. [This implies an action on JC to produce such a definition, but no action was explicitly agreed.] **JC**

§**4 — The sausage machine.** JC queried whether the generation of the header would really be delayed until everything else was complete. LB agreed that the process of header generation was actually incremental, with more information being added at each stage of processing, but that it had been convenient to present it as if it were a separate step.

§**5.1 — Converters.** JC pointed out that Lancaster's model corpus, which contains considerably less than one million words, should be incorporated into the BNC. The group agreed that production of a converter to satisfy this special case should be undertaken at the appropriate time.

There was some discussion of the upper limit on the time allowed for writing conversion software — potentially eight weeks per million words. GB pointed out that experience to date suggested that useful software could be produced in a much shorter time in many cases. Agreeing that the suggested figure was acceptable provided that common sense was used in its application, the group did not suggest an amendment.

§**5.3 — Semantic checking.** After discussion, the group agreed that the proposal was too prescriptive, and that more latitude should be allowed to those making the check in deciding what to check, provided that at least some fixed percentage of each text is examined. (Figures of 5 – 10% were suggested.) **LB**

§**6.1 — Required elements.** Referring specifically to running heads, pull quotes and captions which, while they appear in electronic editions of the Guardian, are difficult automatically to recognize as such, JC queried whether it was practical to include `<caption>` and `<head>` in the *required* category. LB and DD replied that, given access to printed original copies of the material, it should be possible to add such tagging by hand or with machine assistance within the time allotted for the semantic check. This would bring the proportion of such features correctly tagged up to an acceptably high (over 90%) level, although it would be unrealistic to expect 100% correct tagging of all elements. LB believed that Lancaster wanted `<head>` and `<caption>` in the *required* category, as the quality of the information produced by CLAWS would be degraded unless it could distinguish these features from running text. It was agreed to canvass Lancaster's view and incorporate it into the revision of TGCW27. **DD**

SC, noting that certain *required* features were not marked in the Longman Lancaster Corpus materials that Longman expected to contribute to the BNC, wondered whether the required level at markup could be added within the time constraints proposed for the semantic check. Without making any commitment, OUCS considered it likely that it could, particularly if the features were relatively large (`<div`*n*`>`s, for example). (see also §**??**.)

After discussion, it was agreed that `<div2>` and `<div3>` should be moved into the *recommended* category, primarily because of the difficulty of reliably identifying such features in relatively unstructured source material such as magazines, or novels where some chapters are interrupted by lines of asterisks. As a consequence, `<head>`, while it remains in the *required* category, **LB**

is not required where headings appear before unmarked document subdivisions. (Indeed, the CDIF DTD only permits a <head> or <head>s immediately following a <div*n*> tag, so it is not possible to tag such headings.)

JC queried the meaning of "text" as a tag and in the prose describing <div0>. LB agreed that the latter should really read "textual unit". LB gave an explanation of the use of <div0> in grouping together short textual units having similar subject matter or some other common feature which set them apart from other groups of textual units included in the same composite <text>. An example might be a text which contained a number of magazine cuttings relating to cars, where <div0> tags were used to group the cuttings into those related to Alpha Romeos and those related to Aston Martins. **LB**

§**6.2 — Recommended elements** As noted above, <div2> and <div3> are now in this category.

LB pointed out that there was a suggestion implicit in the proposal that either all elements of a particular *recommended* class should be identified in a text, or none should; it would not be acceptable to tag some such elements in a text, but leave others untagged. The group was unhappy with this, agreeing that while some types of research required texts in which each and every element of a particular type was identified, other studies might merely require to excerpt a number of instances of a particular type of feature for detailed examination, without the need for complete coverage. Ultimately it was agreed that, for *recommended* and *optional* tags, the header of each text should list whether they had been applied, and, if they had, whether the coverage was intended to be complete or partial. The level of coverage can be assessed when the semantic check is applied to the text. **LB**

3 The revised CDIF DTD

LB introduced TGCW28, which defines all CDIF elements and entities with the exception of those related to text and corpus headers. (This material is to be added in the near future.) In response to a question from JC, LB said that it would be desirable for the tags in CDIF-like material received from OUP to have attributes and their values specified where appropriate. **LB**

DD gave a brief overview of TGCW25, which lists the entities used in CDIF to replace characters which cannot be represented reliably or at all as single codes in seven-bit coded character sets. It was agreed that the only characters to be allowed in CDIF content (as opposed to markup) should be those which are part of the ISO 646 invariant subset. This makes it very likely that the printed or displayed representation of the content of corpus texts will be unaffected by the particular code set in use on the printer or display, and so aids interchange of corpus texts.

There were reservations about the proposal that the normalized begin- and end-quote characters (` and " respectively) used in *Freelancers* be replaced by the entities &bquo; and &equo;. OUCS agreed to give the matter more consideration before possibly re-presenting it. **OUCS**

LB asked the group whether it was appropriate to copy the CDIF DTD to people outside the project who expressed an interest. There was no objection to this. (See also §**??**.)

4 Report on materials received

DD referred briefly to TGCW23, a report on trial corpus texts received from OUP in December. There was no discussion, as changes in CDIF and in processing at OUP mean that the findings of the document are outdated.

DD and GB reported briefly on two million words of material received from OUP on 5th February. A sample (something over 10%) of this is being processed through the "sausage machine" proposed in TGCW27. While there are some differences between the data format and that required by current CDIF, transduction has presented few difficulties: GB has passed a number of files from The Independent through a purpose-written converter program, enabling them to pass a syntactic check either immediately or after minimal hand-editing. These files now move on to the semantic check, which will require access to the original newspapers, possibly at the Bodleian library. DD is still working on the alterations required to make a number of other texts pass the syntactic check, after which a semantic check can be made against original texts loaned by OUP. OUCS agreed that this work, and a report on its findings, would be complete by 4th March. **GB, DD**

GB reported that the only firm conclusion that could be drawn from examination of the seven sample written texts supplied by Longman in November is that, because of their variability, OUCS needs more texts to examine before it can properly assess the amount of effort required to bring them up to the minimum level of markup suggested by TGCW27. SC agreed to provide more texts, both written and spoken, covering contractual and confidentiality issues if possible by adding another schedule to the existing research agreement between Longman and OUCS. If it is possible to provide the whole of the Longman Lancaster Corpus on these terms, this should be done. There may be some delay in providing spoken material, as existing texts need further proof-reading to ensure that they conform to the principles of TGCW21, the transcription guide. **SC**

5 Arrangements for storage of, and access to, original texts

OUCS proposed that the original texts used in the preparation of the BNC should be brought together as a central resource, firstly for use by the corpus developers, and later to allow controlled access by interested users of the BNC. The British Library or OUCS were possible homes for the material.

While agreeing that this was a good idea in principal, both Longman and OUP wished to retain the original texts that they had captured in order that their lexicographers could refer to them if necessary. It might be necessary to purchase duplicate copies if a central repository were to be set up. It was agreed to refer the matter to the Project Committee. OUCS is to draft the proposal. **OUCS**

6 Text selection issues — brief status report

JC announced that Russell Sweeny, a consultant employed by the British Library, had ten days available for BNC-related matters. Initial discussions between OUP and Sweeny, with input from OUCS, had produced useful insights into sources of information on, and means of selection of, books and serial publications for the BNC. The work is proceeding, with a report sometime in March as the expected outcome.

7 Data protection issues — brief status report

Because the BNC project may hold personal information relating to authors (names, domicile, age...), publishers, copyright holders and users (names and contact information), it falls withing the requirements for registration under the Data Protection Act. DD has enquired of the OUCS Data Protection Officer (Keith Moulden) whether the project is covered by Oxford University's existing registration. Keith has referred the issue to the University's DPO, whose initial feeling is that an additional and separate registration is required by the BNC. A formal reply is anticipated in a week or two.

8 Any other business

8.1 ALLC panel

There was a brief discussion of the format of the panel session involving BNC project personnel at the ALLC meeting of 5th–9th April.

8.2 Information paper

It was agreed that there was a need to produce a technical description of the BNC project which could be sent out in response to serious enquiries from potential users, and those involved in similar activities. The written and spoken corpus design specifications, and the CDIF definition were suggested as elements of such a description. LB suggested that the result might appear in the ALLC proceedings, which would appear sometime after the conference. Both GB and Geoff Leech are writing papers which could also be used for this purpose. group members should assess their suitability for this purpose when they have been produced. **GB**

9 Review of agreed actions

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

10 Close

The meeting closed at 17:00.