TCM01
BNC Technical Subcommittee
Minutes for meeting of 30th November, 1992

Dominic Dunlop

8th December, 1992

Present:

RA	Robert (Bob) Allen	Chambers
MB	Michael Bryant	Lancaster
AGB	Gavin Burnage	OUCS
SC	Steve Crowdy	Longman
DD	Dominic Dunlop	OUCS
FK	Frank Keenan	OUP
RW	Ray Woodall	OUP

Note on document numbering These minutes start a new series of numbers for materials related to the Technical Committee. However, the numbering series for task groups A through D will be continue in use for materials which pertain to work previously handled by particular task groups.

1 Opening of meeting

The meeting convened at 13:40. DD agreed to write the minutes; SC chaired the meeting.

There were introductions for Bob Allen, representing Chambers; and for Frank Keenan, who joined the project as a programmer for OUP at the beginning of November.

2 Minutes of previous meetings

As the technical committee takes over the functions of four task groups, the minutes of the last meetings of task groups A, C and D were reviewed. (Task group B has never had an official meeting.)

2.1 Task Group A meeting of 31st July (TGAM20)

OUCS *Provide copies of definitive written permissions documents (TGBP07) on WordPerfect diskette.*

Done.

OUP, OUCS *Press British Library for BLAISE service.*

Done. We now have an account.

SC *Circulate BNCP25.*

Done.

JC *Provide documentation on random text selection procedures.*

Done. Document is TGAP21.

JC *Subject to funds being available from the British Library after taking the costs of any activities involving the National Sound Archive into account, investigate cost and feasibility of getting librarian assistance to classify whole set of* BNC *texts.*

Open. SC reported that the NSA was willing to archive spoken corpus source material at no cost to the project, leaving the field open for further BL-sponsored research involving written materials. OUP to pursue. **RW**

2.2 Task Group C meeting of 29th September (TGCM37)

SC, RW *Eliminate duplicates from Longman and* OUP *material to be submitted for inclusion in the* BNC.

Done. There is only one such text. There was some discussion, without resolution as to whether it should be counted as having been contributed by Longman or OUP. On the basis of material received to date, DD expressed a preference for data from OUP.

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. Report will be produced when written input has been received from John Wells. A preliminary suggestion is that excerpts of recordings of recruits' voices should be sent to dialect experts for precise classification of their dialects. **SC**

LB *Make proposal to Project Committee on storage of original texts.*

Open. **LB**

LB *Investigate why* ALLC *had not requested paper on* BNC.

Done. There was no communications failure: the ALLC did not want a paper.

DD, RW *Discuss use of* `<head>` *and* `<caption>`

Ongoing. One meeting on the topic has taken place; another has been arranged. **DD, RW**

DD, RW *Discuss impact of new* `<sp>` *and* `<stage>` *tags.*

Done. DD and FK have discussed.

SC, LE, RW *Sign off TGCW30, version 1.2.*

Done. No comments were received.

LB *Issue revision of TGCW27.*

Open. **LB**

DD, RW *Liaise on data transfer issues related to* OUCS *and* OUP *databases, and to contents of corpus header.*

Open. **DD, RW**

SC *Review proposals of TGCW34 for corpus header tags to handle data collected on demographic corpus participants.*

Done. The proposals are acceptable.

LB *Merge proposals of TGCW34 into TGCW30.*

Open. **LB**

SC *Investigate means whereby Longman corpus texts may be cast into a form acceptable for inclusion in the* BNC.

Done. Prior to their submission to OUCS, Longman will subject the texts to additional proof-reading, and insert mark-up corresponding at least to the *required* tags of TGCW27.

SC *State what level of "bounces" of material submitted by Longman for inclusion in the* BNC *might be acceptable.*

Closed. This issue may be re-opened after some experience has been gained in processing Longman texts.

SC, RW *Liaise on number of words ultimately available to* BNC *from Longman and* OUP *corpora, taking into account sample size, permissions issues, Britishness of author, etc.*

Open. SC commented that the current refusal by Reed Group companies to grant corpus permissions was likely to cause a bigger loss in availability than any other factor; RW agreed, but said that recent informal discussions had suggested that Reed might yet have a change of heart. **SC, RW**

SC *Deliver several hundred thousand words of spoken text to* OUCS.

Done. Following discussions between Longman and OUCS, OUCS has not processed this text, as it is to be resubmitted with a different file-naming scheme, with small texts having identical (or very similar) headers consolidated into single files, and with each participant described only once, rather than in each in the header of each text in which they participate.

AGB, RW *Liaise on* Ulster Newsletter *and Thompson regional newspapers.*

Done. OUCS has passed some *Newsletter* material to OUP; OUP will pursue Thompson.

DD *Generate sample tape of* BNC *materials for participants at end of 1992.*

Open. DD will canvass participants by telephone on acceptable media, data formats and contents. **DD**

LB *Ask opinion of Project Committee on issue of balance arising from amount of "quality" newspaper material submitted for inclusion in* BNC, *and absence of other national newspaper material.*

Done. For the moment, approximately one million words from each of the *Daily Telegraph*, *Independent* and *Guardian* has been accepted for inclusion in the BNC. OUP has passed a further million words from each of the *Independent* and *Guardian* to OUCS, but this is being held at this stage in processing with the agreement of both parties, in the hope that it can be replaced by material from different sources.

RW *Investigate further sources of machine-readable newspaper material.*

Ongoing. OUP has recently obtained permissions for the use of material from *Today*. **RW**

LB *Ask at Project Committee whether Chambers would like to participate in Task Groups.*

Done. They would — as shown by RA's attendance here.

LE, LB *LE to canvass interest at Lancaster in participation in* SERC JFIT *poster session; LB to follow up on behalf of* OUCS.

Open. **LB, LE**

2.3 Task Group D meeting of 29th September (TGCM13)

OUCS *Send eight million words of corpus material to Lancaster.*

Done.

LE *Draft proposal on Lancaster's handling of apparent errors in received texts.*

Open. **LE**

DD, RW *Discuss means of representing lineation in headings, where this is used in place of punctuation.*

Ongoing. This issue was one of the aspects of data-capture mark-up discussed at a meeting between OUCS and OUP on 7th December; there will be a follow-up meeting on 15th December. **DD, RW**

LB *Update TGCW30 to reflect current state of C5 tag set; recirculate.*

Open. The document has been updated, but has not been recirculated because further amendments are in the works. **LB**

LE *Propose further "portmanteau tags" if shown to be necessary following frequency analysis on corpus texts.*

Open. **LE**

DD *Add additional C6 tags to* CDIF DTD.

Open. I am inclined to postpone this until the changes required by C6 elsewhere in the DTD have been established. **DD**

LE *Circulate revised TGDW11 if shortening of some tag names acceptable to Lancaster.*

Open. **LE**

LE *Circulate revised TGDW11 if C6 extended punctuation tag set not considered necessary by Lancaster.*

Open.

LE *Circulate TGDW12, describing skeleton parser segment types.*

Open. **LE**

LE *Provide skeleton-parsed material to OUCS for evaluation*

Open.

OUCS *Extend* CDIF *to accommodate parse trees*

Open pending completion of previous action. OUCS

LE *Report on status of Corpus Search and Retrieval software.*

Closed. See §4 below.

3 Interim status reports

3.1 Chambers

Chambers has volunteered to collect miscellaneous, unpublished written material for the BNC. RA reported that a detailed proposal, including costings, would be available by Christmas. While the unstructured nature of such material may cause problems, it is anticipated that there will be few problems with copyright. See also §5. **RA**

3.2 Lancaster

MB reported that 13.5 million words had been received to date from OUCS, and that about two million had been word-class tagged and then post-edited by grammarians. While, currently, every new text presents new problems, MB is fairly confident that 100 million words can be processed within the project time scale.

There was some discussion of Lancaster's procedures, which, except for a few fully post-edited texts, involve the post-editing of a one hundred segment sample from each text. MB and DD will liaise on a means of identifying post-edited segments in order that they can be distinguished from the remainder of a text[1]. **MB, DD**

3.3 Longman

SC reported that a Longman employee is currently "on the road", collecting context-governed material from a variety of points in eastern England and Scotland. 140 hours of recordings (perhaps 1.5 million words) have been made to date. Contact has been made with many regional sound archives, which will be able to provide, for example, legal proceedings and oral history material. A maximum of half a million words will be collected from this source.

In order to ease the bottleneck caused by using two DAT machines both for editing and for dubbing, more machines have been purchased. These are domestic units, and cannot record at the 44.1kHz sampling frequency used to date. (They record at 48kHz, so as to prevent digital copying of CDs, which use 44.1 kHz). This is not seen as a problem, as the machines are used to transcribe material recorded on Compact Cassettes, and either 44.1 or 48kHz is more than adequate to this purpose. The project had originally used 44.1kHz with a view to direct mastering of CDs for publication, but, given the many hundreds of hours of material captured, it seems likely that published material would be resampled at a lower frequency and/or compressed.

The throughput of the transcribers is satisfactory; however, correct marking of overlap is presenting problems: see further §6.

3.4 OUCS

DD reported that slightly over 25 million words had been received from OUP; of these, 13.5 million had been checked and forwarded to Lancaster. Throughput was currently running at about one million words per week. This was

[1] A method has since been agreed — DD

satisfactory, but would need to be increased. Syntactic and semantic checking of material from periodicals and ephemera was proving costly: individual throughput on such material was half that which could be achieved on material from books. OUCS and OUP will meet to discuss changes in mark-up and in instructions to keyboarders in an attempt to resolve this problem. **DD, RW**

Slightly under half a million words of spoken and slightly under a million words of written material had been received from Longman. However, in consultation with Longman, it had been agreed that this should be treated as test data, and not processed for forwarding to Lancaster; it would be resubmitted to OUCS in another form for processing. (See comments in §2.2 above.)

3.5 OUP

RW reported that the obtaining of permissions had been identified as a bottleneck and that, after devoting considerable effort to the problem, the situation was improving. Merely getting a response from some permissions holders was very difficult. DD suggested that a telephone canvassing company might be contracted to call rights holders in order at least to establish the name of the person that OUP management should contact with regards to rights. Use of a contractor would cost money, however, and other options should be pursued before resorting to this one.

In the past it has proved difficult to establish the "Britishness" of the author or authors of a work before moving on to expensive data capture — not least because the concept of Britishness is ill-defined. OUCS has "bounced" a number of texts after establishing from information on dust-sheets, in introductory material or elsewhere, that an author has little or no connection with Britain. OUP is now applying such checks, and the number of bounces is much reduced. (Although not discussed at the meeting, I will take an action to provide an outline of the tests we apply, and attempt to offer rationale.) **DD**

Additional scanner operators and keyboarders have been recruited with the aim of increasing data capture to 6.5 million words per quarter. 5.5 million are expected this quarter.

Popular newspaper material will be obtained from *Today* — see §2.2.

4 Input, Search and Retrieval Package (IS&RP)

Production of the IS&RP is Lancaster's responsibility. SC kicked off the discussion by raising three general concerns:

- Why are we doing this at all?
- Who are we doing it for?
- Will it compete with commercial products from Longman and OUP?

To these, one might add a fourth:

- What does "input" (the word that appears in the Collaboration Agreement) mean in this context?

MB suggested that considerations of the origins of the IS&RP proposal might illuminate SC's queries. The software was wanted by the academic participants. Lancaster's vision was that it should provide a demonstration to corpus *users* (as opposed to corpus builders) of what could be done, rather than attempt to provide for all the possible needs of users. The demonstration would centre on the needs of linguists, as this was Lancaster's primary area of interest.

That said, the current project plan milestones were unattainable because, as the person charged with the production of the package, MB was finding himself spending his time on unforeseen issues concerned with corpus processing — issues he judged more pressing, and more important to the overall success of the project than the IS&RP. He suggested that a "bottom-up" approach, addressing the most desirable aspects of the wish-list that was PCW32, would be the most effective way to produce a worthwhile deliverable.

AGB asked whether, given the resourcing problems that MB had identified, other sources of funding should be investigated. It was agreed that this was an option, but no decision was made to pursue it.

AGB also suggested that the IS&RP should build on existing software used by linguists and others, rather than re-implementing. MB foresaw problems in this because

- It would take time to evaluate the applicability and usefulness of existing software;
- Existing software, such as WordCruncher, is not SGML-aware; and
- Lancaster's experience was that its own existing software would not scale to handle the very large volume of data in the BNC.

The commercial partners were asked about their attitude to contributing software to the IS&RP. SC said that Longman was developing its own OS/2-based tools, and would not be contributing these to the project. RW did not know OUP's position, and would attempt to establish what it was. For Chambers, RA stated that the company had no software of its own, and consequently was interested in using the IS&RP. **RW**

Discussion then moved to the manner of implementation of the IS&RP, the features it should offer, and the types of computer on which it should run. MB will amend update PCW32 to reflect the discussion: **MB**

Manner of implementation: The package is to have a front end and a back end. The former should be window-based; the latter may consist of line-orientated tools which are given input parameters by, and which pass their output to, the windowing environment. Only a few back-end tools would be provided, along with hooks to which users could attach their own applications.

Lancaster has proposed to write the front end as a C or C++ application for the X Window system. It is the input of user queries though this interface which fulfills the "Input" part of the package name. DD strongly urged that an application builder should be used, as it would be likely to cut development time, and, if the right product were chosen, would produce software which would run in other environments besides UNIX. He will provide a list of such products to Lancaster, bearing in mind that anything that requires end users to purchase a run-time licence is undesirable. **DD**

Features offered: The front end should allow users to select texts (and possibly parts of texts) and the back end processing to which they are subjected. It is not clear whether the source of texts should be local to the system running the front end; accessible via Local-Area Network (for example, a complete or partial copy of the BNC held at the user's site; or accessible via Wide-Area Network (a definitive version of the BNC held at OUCS or some other archive site). OUCS has been planning to produce tools to allow users to select and fetch corpus texts from archives; MB and DD will discuss the integration of this function with the IS&RP. **MB, DD**

It was agreed that a central feature of the back end should be an indexing facility, and that this should be able to build, as well as to use, indexes. This would allow users to construct customized indexes for corpus subsets of their own choosing. According to the project plan, there is no requirement that an index is provided for the BNC as a whole, alongside the texts, although this would be desirable.

DD suggested that the community of potential users, perhaps as represented by subscribers to `corpus` and other mail lists, should be canvassed for their "wish list" of features in the IS&RP. As implementors, Lancaster should ask the question. MB agreed to this, but pointed out the need for circumspection if the expectations of users were not falsely to be raised. **MB**

FK will try to locate such a wish list, which he remembers seeing in SALT Club proceedings. RA will provide Chambers' wish list. **FK** **RA**

Machines supported: Lancaster's proposal addresses only the UNIX environment. OUCS believes that this considerably limits the community of potential users, since many researchers in the humanities prefer to use — or have no option but to use — PCs or Macintoshes. Other things being equal, it would be desirable to aim for a portable package. The use of an application builder would aid front-end portability, and line-orientated back-end programs can, if written with care, be made easily portable. However, the limitations and idiosyncracies of particular environments — for example, limited memory space on MS-DOS, and the limited processing power and disk space of many PCs — could result in a portable package being less capable than one developed exclusively for fast, high-capacity systems running UNIX. The balance between these factors requires further consideration.

5 Chambers miscellaneous, unpublished data capture plans

As stated under §3.1, a firm plan will be produced by Christmas. DD queried the format in which material would be delivered to OUCS. RA replied that OUP's transfer format would be used.

6 Longman spoken data capture plans

SC had been asked to report on means by which Longman could achieve the target of ten million words of spoken material for the BNC. He canvassed the views of those present on nine options, each of which would yield an incremental increase in the amount of data captured. A written report on the topic will shortly be complete; the following should be regarded as a preview. **SC**

1. **Use freelancers provided with DAT recorders for editing and dubbing**, so removing a processing bottleneck. This has already been put into effect. (See also comments under §3.3.)

2. **Chance balance from 50–50 to 60% context-governed, 40% demographic**, as transcription of context-governed material has proved less costly and problematic than that of demographic. As approximately 4.5 million words of demographic material have already been recorded, about half a million would remain untranscribed if this option were taken up.

3. **Use material from sound archives**, so eliminating time and cost of recording some context-governed material. Approaches have been made to archives, but no material has yet been used.

4. **Use existing transcribed material** — for example, those made for production or legal purposes by broadcasters. It remains to be seen whether such material is available in a format which accurately transcribes what has been said (including false starts and so on).

5. **Increase amount of broadcast material from 10 to 20%**. The cost of recording such material is low, although obtaining permission to use it in the BNC can present problems.

6. **Increase maximum sample size from 10,000 to 40,000 words**. The original limit was arrived at somewhat arbitrarily by estimating the capacity of a single cassette. Experience has shown that considerably larger samples can be captured in context-governed situations.

7. **Reduce the number of regional sampling points for context-governed material**, so easing logistics and cutting expense. (The reduction would only be relative to Longman's internal plans; the corpus design does not specify the number of sampling points, only their regional distribution.)

8. **Group short conversations involving the same participants under a single header**, so cutting the overhead of creating large headers for many tiny files. It has already been agreed with OUCS that this is acceptable.

9. **Reduce or discontinue the marking of overlap**, as transcribers' output contains many errors, and these are expensive to correct.

These proposals were generally acceptable, with two exceptions. Firstly, changing the balance of the corpus (option 2) was seen as something to be avoided if possible. Secondly, any reduction in the marking of overlap (option 9) was unacceptable. MB pointed out that the output of CLAWS would be of little use unless utterances containing complete sentences were captured as such, even where overlapped. This implied attention to overlap, which, in turn, suggested that it should be marked. SC agreed, saying that he thought that, taken together, the other options would bring the size of the spoken corpus up to ten million words without taking the drastic step of ignoring overlap.

DD proposed that tools, perhaps derived from the `reformat` and `check` programs (see TGCW32) provided by OUCS, should be put in the hands of transcribers, allowing them to correct their own errors, rather than leaving clean-up to a later and more costly proof-reading process. Again, SC agreed, and asked whether OUCS could help in the adaptation of the tools. DD regretted that OUCS had no spare resources to do this.

7 Any other business

AGB reported that he and Roger Garside of Lancaster had submitted a proposal for a joint paper to be read at the 1993 AHC-ALLC meeting, to be held in Georgetown (Washington DC) in June.

8 Review of agreed actions

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

9 Date of next meeting

No date was agreed.

10 Close

The meeting closed at 16:45.