Minutes of BNC Task Group C Meeting, 5 June 1991

Lou Burnard

July 31, 2009

1 Present

DD (chair), GB, JHC, LB, SC (in part)

2 Procedural matters

DD's agenda was agreed. This being the first meeting there were no outstanding items. The document list for the Task Group was reviewed. It currently contained the following items:

TGCW01 Burnard *Markup scheme for the BNC*, 25 April 91

TGCW02 Leech *Basic grammatical tagset initial proposal* and *Penn Treebank Tagset*

TGCW03 CPH Appendix A 13 Jan 89 (Longman's original proposals for encoding corpus materials)

TGCW04 Clear *Markup for the Oxford Pilot Corpus*

TGCW05 Burnage *Database Design Specification*

TGCW06 Dunlop *Text Submission Guidelines*

TGCW07 Dunlop *Encoding the Oxford Milton*

It was believed that TGCW03 was the most recently received version of the tags used by the Longmans texts to be included in the corpus. (SC later confirmed this).

3 Discussion of TGCW01

The main item was the proposed CDIF tagset, as defined in document TGCW01. LB apologised for not having revised the draft in light of comments received to date. There had been an internal discussion within OUCS, and two sets of comments had been received from Lancaster.

The alphabetical list of tags in TGCW01 was gone through in some detail. An attempt was made to reach consensus as to whether distinguishing each feature in CDIF was Essential Desirable Nice or Undesirable, and some attempt at predicting whether making such distinctions automatically would be Easy Tricky or Impossible. The results are listed below, and will also be incorporated in the next revision of TGCW01.

The following general points arose during the discussion:

- CDIF defined an interchange format, which need not necessarily be the same as either a data capture format, appropriate to the needs of those entering or editing texts, or a data processing format, appropriate to local software. LB said that while making the processing format identical to CDIF would obviously simplify matters at any site, there was certainly no need for data to be captured in CDIF. OUCS would attempt to convert data capture formats to CDIF where this could be done automatically, and would refer any problematic material to OUP for consideration.

- Textual features which could not be identified automatically with any degree of reliability, other than those which by definition had to be entered manually (such as the editorial tags **<corr>**, **<add>** etc.) could not be mandated for CDIF, for obvious practical reasons. There was also considerable dispute about textual features thought to be ill-defined or inherently controversial, notably lists.

- The emphasis on descriptive, as opposed to presentational, markup was likely to cause most problems, both in converting from existing material such as the Oxford Pilot Corpus and the Longman material.

3.1 Alphabetical list of CDIF features

In the following list, tags are listed alphabetically, with a cross reference to the section in the TEI Guidelines where the corresponding feature is defined.

abbr [P1 abbrev, 5.3.7]. Agreed to be too tricky to identify and of very little use.

add [P1 add, 5.4]. Agreed to be desirable and (by definition) feasible.

address [P1 address, 7.5.3]. Somewhere between Nice and Desirable, but tricky to identify. JHC noted that existing texts in the pilot corpus contained many addresses.

back [P1 back, 5.2.5]. Agreed to be Essential and Easy.

body [P1 body, 5.2.4]. Agreed to be Essential and Easy.

citn [P1 citn, 5.5]. There was some discussion of the distinction between this and **<q>**. It was agreed that embedded citations and references should be marked using the tag, but that internal components (such as author, title etc.) would not be distinguished. Felt to be tricky.

corr [P1 corr, 5.4]. Agreed to be desirable and (by definition) feasible.

date [P1 date, 5.3.11]. Agreed to be Nice but Tricky.

div1 [P1 div1, 5.2.4]. Agreed to be Essential and Easy.

emph [P1 emph, 5.3.2]. Agreed to be highly problematic to identify and of little use.

enum [P1 enum, 5.3.8]. Agreed that it was Desirable to mark these whether they appeared within a formal list or as a floating reference to a list item. Much controversy as to whether other parts of a list structure should (or could) be formally identified.

figure [P1 figure, 5.9]. Felt to be Essential. No consensus as to feasibility.

foreign [P1 foreign, 5.3.4]. No consensus as to importance. Generally felt to be impossible to identify.

front [P1 front, 5.2.3]. Agreed to be Essential and Easy.

head [P1 head, 5.2.4]. Agreed to be Essential, some doubt as to feasibility.

header [P1 tei.header, 4]. Agreed to be Essential and Easy.

hi [P1 highlighted, 5.3.2]. Easy to tag, if typographically marked. No consensus as to usefulness. Much discussion as to need for both this and **<q.mark>**, qv.

in.quot [P1 in.quot, 5.3.3]. Agreed to be of little use and Impossible to mark auomatically.

item item [P1 item, 5.3.8]. No consensus. LB felt that it might be possible to tag automatically by the presence of **<enum>**s; JHC disagreed, and felt that the concept was too ill defined to be of use.

l l [P1 l, 7.3.1]. Agreed to be Essential and Easy.

list [P1 list, 5.3.8]. Agreed (with some reservations from JHC) to be Essential. No consensus as to feasibility.

name [P1 propname, 5.3.6]. Agreed to be Desirable but Tricky.

note [P1 note, 5.3.9]. Agreed to be Desirable/Essential, and Easy to mark.

number [P1 num, 5.3.11]. Some doubt as to utility, but Easy. In spoken texts, felt to be essential, as the intention was to normalise.

p [P1 p, 5.3.1]. Agreed to be Desirable/Essential and probably Easy to detect.

point [P1 milestone, 5.6.4]. Agreed to be Desirable/Nice, only feasible if reference points were already marked in source material.

q [P1 q, 5.3.3]. Agreed to be Desirable, but felt to be Tricky

q.mark [P1 q.mark, 5.3.3]. Disagreement as to the need for this feature as well as **<hi>**. Agreed that it was Easy, though differentiating from **<q>** was probably very Tricky.

s [P1 s, 5.8]. Agreed to be Essential; no consensus as to feasibility. It was suggested that segmentation might be better left to the Lancaster Parser.

text [P1 text, 2.4]. Agreed to be Essential and Easy. LB noted that more attention was needed to the implications of using samples rather than whole texts in this context: sample boundaries leading to textual discontinuity would need to be marked in some way.

turn [No P1 equivalent]. Agreed to be Essential and Easy.

w [No P1 equivalent]. Agreed to be Essential at some level, but tricky in that the Lancaster Parser might tokenise the text differently from the input.

LB reported that Michael Sperberg-McQueen had been assigned to the National Corpus project as TEI Consultant and would be reviewing CDIF and advising on TEI-compatibility at the forthcoming TEI Workshop (1-5 July).

Time precluded discussion of the other documents tabled at the meeting. No date was fixed for the next meeting.