TGC W27
BNC Acceptance Procedures
Draft OUCS Proposals

Lou Burnard

15 January 1992, revised 6 March

1 Introduction

Discussion in Task Group C and at the Project Committee, indicates some urgency in reaching formal agreement between BNC participants concerning acceptance procedures and conformance thresholds for texts entering the Corpus 'sausage machine'. This note sets out OUCS proposals in this respect, as revised following discussion with Task Group C. It represents the consensus reached by that group, and is being presented for approval by the Project Committee.

The job of reaching an agreement seems to have the following components:

- agreement as to the content and structure of CDIF, and in particular which of the textual features it distinguishes are mandatory for acceptance purposes
- definition of the format or formats in which texts will be supplied to OUCS by participants, and the relationship between that format and CDIF.
- agreement as to the procedure OUCS should follow in validating the materials once they have been converted, and the thresholds at which materials should be excluded from the corpus

This document is mostly concerned with the third of these, but the following general remarks on the other two may be useful to place the rest in context.

2 CDIF

This is the target format for the whole corpus, spoken and written. A working paper (TGCW25) documenting it in detail is currently being produced. Its basic content has already been largely agreed to by all participants, and is summarised in section 4 below. This list has been expanded beyond that previously agreed by TGC to include tags needed for spoken text, but is otherwise largely unchanged.

3 Delivery formats

Participants may of course, as OUP has, elect to supply material in CDIF directly, but this is not essential, provided that definition of an automatic conversion procedure is feasible and cost-effective, as (for example) with the Longman's spoken material. It seems likely that we will need to provide a number of such procedures for texts coming to us from different routes, as not all sources of corpus material may wish to convert to CDIF themselves. This is obviously the case with material in pre-existing electronic form. OUCS is willing and able to do this, provided that we receive a full and accurate definition of the format used, together with a sufficiently large sample to test our conversion procedures on. If, in our estimation, any set of materials provided is so heterogenous that no single procedure or set of procedures could achieve the target rates of throughput and accuracy (see below), we will propose that the material (or some other exemplars satisfying the same selection criteria) be rekeyed.

4 The sausage machine

The following processing steps are envisaged for each text received at OUCS:

- Non-CDIF texts are first automatically converted to CDIF. Texts for which no converter exists are not accepted into the machine
- Syntactic accuracy of the markup is checked, using an industrial strength SGML parser
- Semantic accuracy of the markup is checked at regular intervals throughout the text, using techniques outlined below
- Texts which do not pass both sets of checks will be referred to the supplier with a suggestion that the material (or some other exemplars satisfying the same selection criteria) be re-keyed.
- Texts which do pass both sets of checks will be batched up and transmitted to Lancaster for enrichment
- Texts received from Lancaster will be automatically checked for syntactic accuracy a second time. We do not anticipate a need for further semantic checking.
- Inclusion of texts into the project database, generation of standard header etc. This process is independent of the others and thus can be carried out in parallel with them.

5 Proposed Thresholds

5.1 Converters

In general, we will develop customised software only for formats in which large amounts (i.e. more than 1 million words) of texts are anticipated and for which we have been provided with a full and accurate description. [1] If, in our opinion, provision of such software would require investing more than 6 person/hours per 20,000 words, we will not undertake it. Conversion of small quantities of text using combinations of ad hoc tools will be undertaken on a best-endeavours basis, and only where it does not lead to our overall throughput of material falling below the target levels implied by project milestones.

5.2 Syntactic checking

All texts in the corpus must parse correctly. We will do our best to fix any systematic or sporadic tagging errors causing SGML parser error messages, provided that doing so does not take up more than 1 person hour per 20,000 words.

5.3 Semantic checking

Only syntactically valid texts will be checked for semantic accuracy. For this purpose, we will need copies of the original source material against which to carry out spot-checks of the encoded text. The checks will be carried out visually against the start and end of each text and at several randomly chosen places within the text. Between 5 and 10 percent by bulk of the texts should be checked against the original in this way in order to determine whether, in our opinion, all the textual features which should have been tagged (i.e. those mandated by CDIF) have in fact been tagged. Any obvious typos, missing words etc will also be noted.

We will do our best to fix all such errors (throughout the text, not just in the pages checked), but only where we are confident that doing so will not take up more than 6 person hours per 20,000 words. While achieving 100% accuracy is recognised as impossible, our intention is to correctly identify and tag over 90% of occurrences of features distinguished by the markup scheme.

5.4 Overall limits

The threshholds quoted above are subject to the further constraint that we do not have the resources to spend more than 40 person hours a week on correcting syntactic or semantic errors in texts submitted for inclusion in the corpus.

[1] An exception already agreed is the Lancaster "model corpus"

6 CDIF summary

6.1 Required

The following textual features are mandated by CDIF. If retained in the text as captured, they must be marked up, using the tags shown.

cdif Tags a single conformant CDIF text (SW)

div Any arbitrary grouping of utterances within a spoken text (S)

div0 A group of **<div1>** elements within a written text, which have been combined for convenience of handling (W)

div1 Major subdivision of a written text, e.g. chapter

head A title or heading occurring at the start of a **<div1>** or **<div0>** element (W)

header Bibliographic and descriptive information about a text supplied for BNC indexing purposes (the format and exact content of this information has yet to be agreed) (SW)

item An item in a list (mandatory within list) (W)

l A line of verse (mandatory within poem) (SW)

note A footnote or sidenote in a written text, not forming part of the main text (may be simply deleted from source) (W)

p A paragraph in a written text (W)

ptr empty tag pointing from one part of a text to some other element: used to align parts of a spoken text with a timeline representing overlap (S)

stext An individual spoken text (S)

text An individual written text (W)

u An utterance by a single speaker (S)

6.2 Recommended

Distinguishing the following textual features is regarded as highly desirable. If present in a text, they should be marked up using the tags shown, provided that this can been done throughout a text in a reliable and consistent manner. Some indication should be provided as to which tags in this category have been supplied in a given text, and whether coverage is intended to be complete or partial.

add An editorial addition, supplying for example a word missed out unintentionally during transcription (SW)

caption (1) A heading, title etc. attached to a picture or diagram, usually with deictic content (2) a 'pull quote' or other text about or extracted from a text and superimposed upon it to draw attention to it (may be simply be deleted from source) (W)

del An editorial deletion; in spoken texts, particularly where words identifying persons or places have been removed in transcription (SW)

div2 A further subdivision of a written text, entirely contained within a div1, e.g. section (W)

div3 A further subdivision of a written text, entirely contained within a div2, e.g. subsection (W)

head A title or heading prefixed to a **<div2>** or **<div3>** (W)

hi A passage of written text which is typographically highlighted for example by italics or bold, where the reason for this cannot be expressed by other tags (may be simply deleted from source) (W)

label An enumerator or other label attached to a list item, or appearing freely within a text (may be simply deleted from source) (SW)

list A collection of distinct items flagged as such by special layout in written texts, often functioning as a single syntactic unit (may be simply deleted from source) (W)

pause A marked pause during or between utterances in a spoken text (S)

pb Marks the start of a new page in the original source (may be simply deleted from source) (W)

poem A poem or extract from one, embedded or quoted within a text (may be simply deleted from source) (SW)

reg Any editorial regularisation, whether to correct a word or phrase mistranscribed or mis-spelled, or to normalise variant spellings. (SW)

shift A marked change in voice quality (S)

sic A word or phrase which has not been regularised, but which is in doubt, for example a spoken word which the transcribers cannot recognise. (SW)

trunc A word or phrase which has been truncated during speech (S)

unclear A point in a spoken text at which it is unclear what is happening, e.g. who is speaking or what is being said (S)

vocal A non-linguistic but communicative noise made by one of the participants in a spoken text (S)

6.3 Optional

Distinguishing the following textual features is entirely optional at the corpus acquisition stage. They are provided for the convenience of participants wishing to preserve features already encoded in texts.

abbrev Any acronym or abbreviation. In spoken texts acronyms are spelled out as they are pronounced, but need not be tagged as such (SW)

back Matter not forming part of a text but appended to it in an appendix or similar (may be simply deleted from source) (W)

citn A bibliographic citation, containing possibly an author, title, page reference etc. (W)

date A calendar date in any format (SW)

epigraph A quotation or dedication prefixed to some division of a written text (may be simply deleted from source) (W)

event A non-communicative event (e.g. a door slamming) occurring during a conversation regarded as worthy of note. (S)

front Material prefixed to but not forming part of a written text (may be simply deleted from source) (W)

marked A word or phrase regarded as marked, for example as non-English, technical, archaic, regional etc. (may be simply deleted from source) (SW)

propname Proper name of a person, place or institution (SW)

q A quotation, either embedded or displayed; also, any representation within a written text of spoken language (e.g. dialogue) (W)

salute A formulaic greeting, appearing at the start or end of some unit of a text. (W)

title The title of a book, song or similar bibliographic entity, either within a **<citn>** or cited elsewhere in a written or spoken text (SW)

6.4 Generated

Markup of the following features will be automatically generated during the corpus enrichment process at Lancaster.

s A segment of text corresponding to the CLAWS segmentation scheme (SW)

word class codes These will be converted to pointers linked to a TEI-conformant *feature structure declaration*. Further details to be supplied (SW)