TGCW06: Text Submission Guidelines: Draft for comment

Dominic Dunlop

29th May, 1991

1 Sources of Text

The British National Corpus project will be receiving text from the following sources:

- Existing on-line or archived computer data;
- Text captured using optical character recognition equipment;
- Rekeyed texts; and
- Transcribed texts (from spoken corpus source tapes and broadcast material).

On-line data, broken down as 20% from existing corpora and 30% from other sources, is expected to account for 50% of the corpus — that is, 50 million words; newly-captured text to make up the remaining 50%.

Text from all sources must be marked up in as consistent a manner, and to as consistent a level, as possible using the BNC markup, CDIF (Corpus Document Interchange Format) defined in [**?**]. It may be possible to mark up text from some sources using the BNC markup, or a useful subset thereof, at the time that it is captured. Alternatively, captured text may be marked up using some alternative scheme — for example, that defined in [**?**, annex] — which can relatively easily be mapped into BNC markup. Section 3 explores these issues for each particular source of text.

2 Goals for OUCS Processing Procedures

As set out in [**?**], texts received by OUCS will pass through the following processing stages:

- Reading source

- Making files from each text
- Conversion to CDIF
- CDIF consistency check
- Syntactic tagging (at Lancaster)
- Final CDIF consistency check
- Accession to the Corpus

It may be that, in particular cases, particular steps will be empty: for example, if it proves feasible to apply CDIF tagging to spoken corpus material as it is transcribed, there is no need for a subsequent conversion process. Instead, the text can move straight to the consistency check. (That is, a check that it conforms to the CDIF DTD.) Similarly, the source reading step may be empty for any texts which arrive by network file transfer. In general, however, all texts will be subject to processing at each stage.

The expected volume of texts to be processed for accession to the hundred-million word Corpus is high. If sampling were to be used[1] to set a maximum text size of 40,000 words, the number of texts would be something over 2,500. (Some texts would be shorter than 40,000 words.) If a large number of small texts — for example, individual business letters or items of ephemera — were included, bringing the average word count down to 10,000, the Corpus would ultimately contain 10,000 texts. As the project timetable dictates that accession at OUCS will take place over a two-year period from October, 1991 to September, 1993, it will be necessary to process between 25 and 100 texts per week on average through each of the steps listed above.

Clearly, given this volume of material, it will not be possible to lavish much individual attention on any one text. It will certainly not be feasible to use an approach which has been adopted on some smaller corpora; that is, to 'hand craft' each text in order to bring it into line with a target encoding. Instead, the project must aim to develop procedures capable of processing particular classes of text with as little human intervention as possible.

In developing a procedure, the first few texts that it processes will be used as test data for prototype versions, and will therefore receive individual attention. As the procedure is successively refined, it should be possible to process further texts with less and less intervention. Ultimately, the procedure should identify and flag only a small number of situations that it cannot handle, otherwise running without human input. (It seems desirable, however, to perform some type of quality check on texts which have been processed without complaint by a procedure, so as to be assured that the procedure is working correctly.)

[1] Sampling complicates the application of structural tags, and so would increase OUCS' workload.

The project database will document the procedures used in bringing each received text from its intial state to a state suitable for accession to the Corpus. This information will be useful both in determining how to process new but similar texts, and, should a Corpus text be lost, in reconstructing it. (Archive copies of each text will be made at key processing points in order to guard against such loss; the ability to re-run procedures will be used only as a secondary recovery method.)

3 Processing of Incoming Texts

As section 1 stated, texts will come into OUCS from a variety of sources. These are discussed in the following subsections.

It is important to realise that, by the time OUCS gets its hands on any text, it will always be in computer-readable form. Thus, subsections 2 to 4 describe important special cases of the sources for the computer data discussed in subsection 1.

3.1 Existing computer data

In handling material of this type, OUCS faces four levels of variability:

- The delivery medium (magnetic tape, data cartridge, diskette, network file transfer, etc.)
- The delivery format (labelled tape, VMS BACKUP, UNIX tar, MS-DOS pkarc, etc.)
- The source text encoding (typesetter commands, TEX, WordPerfect, Quark Xpress; ASCII, EBCDIC....)
- The style in which the encoding is applied; for example, the sequence of typesetter commands used to start a new paragraph or to set out a running head.

The project must develop procedures to handle each of these levels. It is anticipated that the most difficult problems will occur at the third level, in mapping from the presentational markup which is typical of current text encoding schemes to the structural markup required for the Corpus. Variability in the means by which particular presentational features are expressed, even within a particular presentational encoding, will further complicate the issue, and make it more difficult to create general-purpose tools capable of (more or less) automatic translation from a particular encoding to CDIF. Experience to date [**?**] suggests that the identification of structural features in presentationally-marked texts is a time-consuming and skilled task. A very rough estimate of the time

required to create a robust procedure capable of giving useful translations of all texts in a given encoding is one man-month.

For these reasons, it seems desirable to limit the number of encodings with which OUCS has to deal when processing texts received as computer data. While the fact that a given text uses a hitherto unseen encoding should not preclude its accession to the Corpus, it might count against it. Serious consideration should be given to rekeying or scanning texts which, although available as computer data, are in a format which is unlikely to be encountered more than once. A possible alternative is that texts available in unusual word-processor or desk-top publishing formats could be converted to 'flat' ASCII data using the software package in question before being submitted to OUCS. (Although this would introduce a serious risk of losing important presentational markup, and so compromising the later conversion to CDIF.)

It would be helpful to OUCS if important[2] source encodings could be identified as soon as possible. This would allow work on translators to be scheduled, and would help in the planning of submission dates for texts using particular encodings. Without such planning, it is likely that work on any given translator will be fragmentary, and that depressing backlogs of as-yet untranslatable texts will build up.

3.2 Text captured using OCR

The KDEM (Kurzweil Data Entry Machine) is (with a little operator assistance) good at identifying and tagging local floating features such as typeface changes, sub- and superscripts. It is bad at identifying larger features, whether presentational — such as text set off in some manner — or structural — for example, paragraphs.

These deficiencies can be corrected at one or more of several processing steps:

- KDEM operators can be trained to add mark-up identifying features of interest;
- Proof-readers can add mark-up after a preliminary version of the text has been delivered by KDEM operators; or
- Software on can intuit the existence of the features and add mark-up as part of the OUCS Corpus accession process.

In the first two of these stages, there is a limit to the amount that KDEM operators or proof-readers can do, given the constraints of time and budget. Indeed, the need for a separate proof-reading stage should be avoided if possible, with marked-up data being passed straight from the KDEM to the OUCS BNC project team.

[2] An important encoding is one used by many source texts, or one used by a small number of texts which are judged to be prime candidates for inclusion in the Corpus.

In the last stage, there is a limit to the amount that a program can safely infer. For example, a test text provided to OUCS by OUP, Danielle Steele's romantic novel *Daddy*, does not have identified paragraph breaks. It may be possible to guess at break locations by searching for short lines which end sentences[3], but such a method is not foolproof. If correct identification of paragraph breaks is judged important in Corpus texts, it must be done either by manual tagging with reference to the original printed text, or by ensuring that the presentational cues necessary for later automatic tagging (vertical white space, first line indent or whatever) are not lost when the text is captured.

Similar arguments can be advanced to cover the tagging of other structural features. The main consideration is that, if the tagging of a feature is judged to be too costly at the data capture stage, sufficient information — whether presentational or otherwise — must be retained in the the text received by OUCS to allow subsequent automatic tagging to be reliable.

3.3 Rekeyed texts

Considerations for rekeyed texts are very similar to those for texts captured using the KDEM — indeed, it seems likely that texts will be rekeyed only when, for whatever reason, they are unsuitable for KDEM capture.

It may be appropriate to use a syntax-driven editor[4] for rekeying, so as to ensure that typists produce output which conforms to a (possibly simplified) version of CDIF[5]. The choice of editor and its configuration so as to provide helpful feedback in the event of data entry errors would be important factors in arriving at a solution which is both acceptable to users and cost-effective.

3.4 Transcribed texts

Given that a suitable set of tags can be developed and agreed, it seems reasonable to expect that the transcribers can create marked-up files which will need little or no further translation by OUCS. Again, the use of a syntax-driven editor would help in the production of texts which conform to CDIF.

4 Further Work

The considerations set out in this document suggest that refinement of the text submission guidelines requires that further work is done in the following areas:

[3]That is, lines which would not fill to the full page width when set in some likely font, and which end with a sentence terminator.

[4]Presumably running on a PC, or perhaps a Macintosh

[5]An example of such a simplification might be to allow the entry of quotation marks, rather than to require the placement of appropriate tags

- Identification of important delivery media, delivery formats and source encodings.
- Development of techniques for translating presentational markup into CDIF.
- Refinement of estimate of development time for new translation procedures.
- Identification of means by which the KDEM or its operators can preserve presentational features required by later tagging procedures.
- Partitioning of markup tasks between KDEM operators, proof-readers and OUCS procedures.
- Investigation of the use of syntax-directed editors.

References

[1] TGCW01: *Markup scheme for the British National Corpus,* Lou Burnard, 25th April, 1991

[2] TGCW04: *Encoding and markup for the Oxford Pilot Corpus,* Jeremy Clear, 17th September, 1990

[3] TGCW05: *BNC text processing and database design: some draft proposals,* Gavin Burnage, 29th May, 1991

[4] TGCW07: *Encoding the Oxford Milton — Interim report,* Dominic Dunlop, 31st May, 1991