CODING METALANGUAGE: ISSUES RAISED IN THE CREATION AND PROCESSING OF SPECIALISED CORPORA

Antoinette Renouf, University of Birmingham, England

In the creation of a computerised corpus of text, it is useful, and sometimes essential, to introduce a measure of coded information as a guide to the interpretation of the linguistic data.

A first and obvious case is in a corpus of spoken language, where the graphic form is only a partial representation of the facts of a communicative event. Where original audio, video or multi-media recordings are available, the basic transcript may usefully be supplemented with prosodic or paralinguistic information. The London-Lund Corpus offers a very full analysis along these lines (Svartvik et aI, 1982).

Another use for codification is where sub-sets of data within a corpus are to be compared statistically by automatic means. Strategically chosen and placed markers serve the dual function of flagging the beginning of each sub-text and of classifying it, so that word counts and other lexical prof1les may be run on individual textual components.

Coding is also called for where the corpus is to be concordanced. It eases considerably the task of interpreting the concordanced output - typically extracts of only one line - by providing further contextual information. Extra-linear coding can be used to identify text source, page reference, and other environmental features. Intra-linear coding helps to articulate the discourse structure, by marking participant change, in spoken dialogue, and various other discourse features. This is especially important in a text corpus of a more fragmented kind, such as one of EFL teaching materials (Renouf, 1987). In Figure 1, a concordanced extract from the Birmingham TEFL Corpus demonstrates the two types of coding:

FIG 1: Concordance Extract for the word form STANDARD (foreshortened lines)

e007 s, children *4 and oaps open:	standard	hours except sun mornings may-s In
e007 k chips and peas *3 of a good e029	standard.	London restaurants some of t in *4
e029 ducation and must have a high	standard	written and spoken english.
e008 ugby and reach a ve!)' high * 3	standard.	many of them have in fact playe
e011tables *3 made by a farmer. a e010 nd	standard	measurement for oil. *0 exercise
e010 i should have a reasonable	standard	of *1 living.' *0 exercise *0 if yo of
e015 climate cities and towns the	standard	living *3 roads people food g of
e017 r advertisement *3 in the new e019 s,	standard	november 14th. you may reme of
e019 s,*0 social security and the	standard	living in this count!)'. *0 prov of
e004 lenging: here requires a high	standard	work. life insurance: *3 an ar

The left-hand codes above simply indicate the EFL book source for each concordance line. It is possible to elaborate the source code, for example to include page and even line reference, and it is sometimes necessary to do so, when the corpus is part of a larger text archive. In a technical sub-corpus, one might need a coding like: TW ENGELE 001, where TW stands for 'Technical Written Corpus', in contrast with other archive components, and ENGELE stands for 'Engineering, sub-class Electrical'. It is, however, important to balance the precision and readability of the coding with the requirement to retain maximal context in the concordance line itself.

The in-text codes shown in Figure 1 mark the type of language in the line, differentiating between metalanguage (*0), constructed spoken text (*1), authentic speech in transcription (*2), constructed written text (*3), and authentic written text (*4). Where different codes occur in the same line, the second marks the start of a new text-type.

Although such embedded coding in text is useful, the process of establishing categories is often problematic. It is straightforward to differentiate between a sports report and a weather forecast, or between one newspaper article and another, because this is familiar territory. But in new areas, and where the categories are to some extent still a matter of intuition, the choice may not be so easy. In such cases, coding often involves the circular dilemma of having to make somewhat ad hoc judgements, where decisions would be better made on the basis of linguistic facts which become available only when the sub-texts have been processed automatically. Biber and Finegan have demonstrated the desirability of an empirical approach to text-type categorization (1986).

Some of the issues that are involved in in-text coding may be illustrated with reference to a specialised corpus of EFL examination papers published by the University of Cambridge Local Examinations Syndicate. The corpus is one of several smaller foci of research interest at Birmingham. The aim here is to separate the language types which make up the corpus with a view to monitoring each of them in various ways, using the same categories as given for the TEFL Corpus above, which were based on intuition and initial observation.

Differentiating between language types 1 to 4 in the examination papers is in itself difficult, although less so than in EFL teaching materials. The chief problem lies in distinguishing 'authentic' and 'constructed' language. There are degrees of authenticity to be accounted for since, in the EFL field, texts are often regarded as authentic which have been edited or abridged. But in this paper, I shall concentrate on language type 5, the 'metalinguistic' category. Metalanguage is that feature of all text which serves to organise the reader's perceptions of the writer's message, inhabiting predominantly the 'interactive plane' of discourse (Sinclair, 1983). In examination papers, it plays an important role in instructing the examinee how to proceed and, since it is not generally taught in language courses, it certainly deserves some research attention. It is also note-worthy in being the only part of the paper that communicates directly with the examinee-reader.

The following extract is taken from pages 6 and 7 of the 'Certificate of Proficiency in English, December 1986' paper, entitled 'Reading Comprehension'; the first paragraph was originally printed in italics, or in bold where words are given in upper case:

==============================

(Page 6)
SECTION B

In this section you will find after each of the passages a number of questions or unfinished statements about the passage, each with four suggested answers or ways of finishing. You must choose the one which you think fits best. ON YOUR ANSWER SHEET, indicate the letter A, B, C or D against the number of each item 26 to 40 for the answer you choose. Give ONE ANSWER only to each question. Read each passage right through before choosing your answers.

FIRST PASSAGE

The decline in traditional religion in the West has not removed the need for men and women to find a deeper meaning behind religion. Why is the world the way it is and how do we, as conscious individuals, fit into the great scheme?

There is a growing feeling that science, especially what is known as the new physics, can provide answers where religion remains vague and faltering. Many people in search of a new meaning to their lives are finding enlightenment in the revolutionary developments at the frontiers of science. Much to the bewilderment of professional scientists, quasi-religious cults are being formed around such unlikely topics as quantum physics, space-time relativity, black holes and the big bang.

How can physics, with its reputation for cold precision and objective materialism, provide such fertile soil for the mystical? The truth is that the spirit of scientific enquiry has undergone a remarkable transformation over the past 50 years. The twin revolutions of the theory of relativity, with its space-warps and time-warps, and the quantum theory, which reveals the shadowy and unsubstantial nature of atoms have demolished the classical image of a clockwork universe slavishly unfolding along a predetermined pathway. Replacing this sterile mechanism is a world full of shifting indeterminism and subtle interactions which have no counterpart in daily experience.

To study the new physics is to embark on a journey of wonderment and paradox, to glimpse the universe in a novel perspective, in which subject and object, mind and matter, force and field, become intertwined. Even the creation of the universe itself has fallen within the province of scientific enquiry.
(etc...)

Page 7 (facing
page 6)

26 The author says people nowadays fmd that traditional religion is
A a form of reassurance.
B inadequate to their needs.
C responding to scientific progress.
D developing in strange ways.

27 Scientists find the new cults bewildering because they are A too
reactionary.
B based on false evidence.
C derived from inappropriate sources.
D too subjective.

28 Which phrase in paragraph 3 suggests that the universe is like a
machine?
A cold precision and objective materialism
B the shadowy and unsubstantial nature of atoms
C slavishly unfolding along a predetermined pathway D
shifting indeterminism and subtle interactions

29 The new physics is exciting because it
A offers a comprehensive explanation of the universe. B
proves the existence of a ruling intelligence.
C incorporates the work of men of genius.
D makes scientific theorising easier to understand.

30 The author of this passage is A a
minister of religion.
B a research scientist.
C a science fiction writer. D a
journalist.

================================

The first stretch of metalanguage in the text extract in Figure 2 is fairly easy to identify as such, revealing as it does some of the linguistic features typically associated with procedural discourse of this kind. In grammatical terms, it is marked verbally by a preponderance of imperatives, sometimes with modal elaboration, and by use of the present simple or future tenses. In lexical terms, it has a fairly restricted repertory, of items associated with the circumstances of the examination event, such as 'choose', 'read', 'answer', 'each passage', 'a number of questions' and 'in this section'. It is also physically distinct from the rest of the text, residing in separate paragraphs, employing different type-faces, and so on. But the metalanguage found later in Figure 2 is rather different, and raises a number of questions in relation to its codification.

The first concerns the degree of analysis which is needed. Metalanguage of one sort or another exists at a number of levels in the text (leaving aside for a moment the 'passage' for interpretation). One might want to begin by making a distinction between 'metatext' and 'metalanguage', by which the former refers to the examination as a physical entity, while the latter denotes the events within the paper. Compare, for example, the heading 'SECTION B' with 'The author says...'. At the same time one might note that the two concepts also overlap, as in the heading 'First Passage', or in the phrase 'item 26 to 40', and that each is multi-layered. Consider the complexity of the following:

30 The author of this passage is
A a minister of religion.
B a research scientist.
C a science fiction writer.
D a journalist.

This piece of text does not fit the criterion for metatext earlier suggested, but it is nevertheless metatextual in requesting a real-world judgement on the authorship of the 'passage' in question.

Then, at the next level of delicacy, and cutting across the distinctions made above, the metalanguage combines at least two major discourse types or functions – informational language and directives – which it might be helpful to record. Compare 'In this section you will find...' with 'You must choose the one you think fits best.' And moving on, within these functional categories there are further distinctions that could be drawn, for example in terms of specificity. Compare the more general 'You must choose the one you think fits best.' with the precision of 'On your answer sheet, indicate the letter A, B, C or D against the number of each item...'.

Within the reading 'passage' presented for interpretation there exists another, separate range of metalinguistic activity. This has necessarily lost its original interactive power, but it may still warrant attention, depending on the purposes to which the corpus is likely to be put.

A second question which arises during the coding of this examination text is where exactly the metalanguage begins, and where it ends. I have said that metalanguage is, in general terms, easy to identify in such text, but closer inspection reveals instances where what appears to be purely metalinguistic is actually an interweaving of various linguistic and metalinguistic strands. In the piece of text labelled '26' in Figure 2, for example, the initial section, which runs:

> 26 The author says...

is metalinguistic, but the remainder, i.e.:

> ...people nowadays find that traditional religion is
> A a form of assurance.
> B inadequate to their needs.
> C responding to scientific progress.
> D developing in strange ways.

is made up of textual paraphrase. Similarly, the section labelled '28' reveals a combination of three kinds of text. It opens meta linguistically, with:

> 28 Which phrase in paragraph 3 suggests that...,

it then moves into textual paraphrase, with:

> ...the universe is like a machine?,

and then on to textual citation, in:

> A cold precision and objective materialism
> B the shadowy and unsubstantial nature of atoms
> C slavishly unfolding along a predetermined pathway
> D shifting indeterminism and subtle interactions

Thirdly, there is a question of whether, and how, to record the paralinguistic features associated with examination paper discourse which also contribute to the management and understanding of the text. These include the sequences of digits and characters, which have a framing function; the organization and spacing of paragraphs; and the choice of type-face. Use of punctuation is also important, as can be demonstrated negatively by the absence of quotation marks around the citations in the section numbered '28' in Figure 2 above. The message provided by these features is important but implicit. In common with the linguistic component of the text, it depends for its interpretability to some extent on the examinee's previous exposure to similar discourse (the associated notion of 'intertextuality' is re-visited in Ventola, 1987). It is therefore, in principle, worth recording.

I have tried, in this short paper, to introduce some of the issues involved in coding specialised corpora, with special reference to the metalinguistic sub-text in a corpus of EFL examination papers. The corpus creator has to establish criteria for inclusion for each class of sub-text, and this really entails the kind of exhaustive analysis that the subsequent computerisation is intended to provide. A solution is to employ a provisional categorisation, and to refine it in the light of automatically-produced lexical profiles, which would, for example, reveal changing patterns of lexical clustering (see Phillips, 1983, on the lexical structure of text). The corpus builder also has to decide on the depth and range of analysis that will be sufficient for future research purposes, which in turn leads to the task of establishing appropriate coding conventions. Again, these processes become less daunting if the opportunity for postediting is available.

References

Biber, D. and E. Finegan. 1986. 'An Initial Typology of English Text Types'. In J. Aarts and W. Meijs (eds.), *Corpus Linguistics, II.* Amsterdam: Rodopi, pp. 19-46.

PhilIips, M. 1983. *Lexical Macrostructure in Science Text.* PhD Thesis, Dept. of English, University of Birmingham.

Renouf, A.J. 1987. 'Corpus Development'. In J. McR. Sinclair (ed.), *Looking Up*. London: Collins, pp. 1-40.

Sinclair, J.McH. 1983. 'Planes of Discourse'. In S.N.A. Rizvil (ed.), *The Two-fold Voice: Essays in Honour of Ramesh Mohan.* Salzburg Studies in English Literature. Salzburg: University of Salzburg.

Svartvik, J. 1982 *et al.* 'Survey of Spoken English: Report on Research 1975-81'. In C. Schaar and J. Svartvik (eds.), *Lund Studies in English,* 63. Lund: CWK Gleerup.

Ventola, E. 1987. 'The Structure of Social Interaction - A Systemic Approach to the Semiotics of Service Encounters'. In R.P. Fawcett (ed.), *The Open University Series.* London: Frances Pinter.

Acknowledgements

I should like to thank The University of Cambridge Local Examinations Syndicate for allowing me to quote from their material, and the author of the text extract quoted, who unfortunately cannot be traced.