BRITISH NATIONAL CORPUS

Task Group D - Corpus Processing

Minutes of First Meeting
University of Lancaster, September 5, 1991

Present: Michael Bryant, Gavin Burnage, Jeremy Clear, Steve Crowdy, Dominic Dunlop, Roger Garside, Geoffrey Leech.

1. BNC Tagset

Papers from Geoffrey Leech proposing a BNC tagset to be used by Lancaster in grammatically tagging BNC texts (TGDW01 and TGDW01 - version of 5th. September 1991) were considered. This tagset is a more coarsely grained version of the Lancaster CLAWS 2A tagset. Discussion centred around a paper tabled by Jeremy Clear (TGDW03) and focussed on the issues of the use of "portmanteau" tags, the interpretation of the tag definitions, the granular consistency of the tagset, linguistic issues (particularly in relation to common noun/proper noun distinctions) and the usefulness of the tags (and their names) for users of the corpus.

It was *agreed* that:

i. the proposed tagset was, in broad terms, appropriate for BNC.

ii. "portmanteau" tags would be used, in a defined set of circumstances, to indicate a low level of confidence in choosing between two candidate tags.

iii. the tag names would be three character sequences with a logical structure, rather than mnemonics.

iv. Geoffrey Leech would prepare a further version of the tagset for circulation (including the SALT group), taking account of the issues which had been raised within the Task Group.

v. In due course Lancaster would prepare detailed user documentation describing the tagset and its implementation.

2. TEI conformancy

There was discussion about SGML markup of tagged text and the particular problems related to the "splitting" of orthographic "words" into separately tagged syntactic units (e.g. enclitics) and the association of "words" in the Lancaster "ditto-tag" scheme (whereby a group of orthographically separate words are effectively assigned a common grammatical tag). A paper from Terry Langendoen (TGDW02) provided background.

It was *agreed* that:

i. the tag names would be devised as legal SGML entity names.

ii. "split" words would not include white-space so that a <w> tag would apply to the whole structure.

iii. "split" words would be presented so that the original orthography would be preserved if the SGML markup were removed.

iv. the ditto-tag scheme would be retained and each possible ditto tag given an entity declaration.

v. OUCS would prepare the entity declarations for tags.

3. Marking Up

OUCS were anxious to clarify mark-up responsiblities for tagged text (TGCN01).

It was *agreed* that:

i. Lancaster would mark segment boundaries (<s> ... </s>) and insert the appropriate entity references as part of the tagging process. Segments would not be nestable.

ii. other mark-up as agreed by TGC would be applied by the text suppliers or OUCS.

4. Date of Next Meeting

To be agreed.

The meeting finished at 4.30 pm.

MEB September 9,1991