British National Corpus

Lancaster University

TGDW14

Selective Linguistic Annotation

Michael Bryant 22-02-93

1 Introduction

1. At the BNC Project Committee meeting held in Harlow on 18th. January 1993 the implications of the technical problems which had been experienced throughout the project and the resulting reduced timescale within which Lancaster would be required to complete its tasks were discussed and priorities agreed. Subsequently the possibility of obtaining an additional £10,000 to help finance the Lancaster operation arose as a result of an initiative from OUP.

2. This document presents Lancaster's current intentions relating to linguistic annotation in the light of the agreed priorities.

2 Proposed Completion Targets

1. In relation to linguistic annotation the minimum completion targets are:-

   (a) 100M words word-class tagged with the C5 tagset ("main" corpus).

   (b) 2M words word-class tagged with the C6 tagset (the "core"corpus).

2. The amount of linguistic post-editing possible will be dependent upon the amount time which needs to be spent in the future on resolving processing problems related to incoming text features and CDIF conformancy issues (which have dominated the processing work until now) and the amount of extra resourcing that might be available. The proposed amount of extra resourcing would be devoted to ensuring the 100% post-editing of the "core" corpus and improvement of the sampling rate of post-editing for the main corpus (from approximately 100 sentences in five texts to 100 sentences in three). In addition there would be the potential for completing some annotation enrichment of part of the core corpus. The type of enrichment proposed is skeleton parsing.

3 Enrichment Annotation - Skeleton Parsing

3.1 Potential Target

The potential target for enrichment annotation is 50,000 words of **skeleton parsed text**.

3.2 Existing Markup Scheme

1. Skeleton parsing is a limited form of parsing which indicates the structure of a sentence in terms of "major constituents", ie, sentence types, predicates, clauses, and major phrase types[1].

2. The existing Lancaster mark-up scheme uses a system of nested brackets with labels to identify the constituent type and further tokens to indicate coordination and discontinuity.

3. The symbols used are:-

   Square brackets, which enclose constituents.

[ ]	Can be nested.

   Labels, identifying the type of constituent. The label appears immediately after the opening bracket and immediately before the closing bracket. Unlabelled brackets are permitted in defined cirumstances to allow the marking of constituents which cannot be analysed in terms of the standard label set.

Fa	Adverbial clause
Fc	Comparative clause
Fn	Noun clause
Fr	Relative clause
G	Genitive
J	Adjective phrase
N	Noun phrase
Nn	Metalinguistic constituent
Nr	Temporal adverbial noun phrase
Nv	Non-temporal adverbial noun phrase
P	Prepositional phrase
S	Compound sentence or direct speech
Si	Interpolated or appended sentence
Tg	-ing clause
Ti	to + infinitive clause
Tn	Past participle clause
V	Verb phrase

   Co-ordination is identified by the symbol '&' following the label for the first conjunct and '+' following the label for subsequent conjuncts.
   Discontinuity (where linked constituents are separated by other constituents) is indicated by placing the symbol '@' after the closing bracket of the first constituent and before the opening bracket of the second constituent of a linked pair.

4. The words of the text would also be annotated with word-class tags.

[1] UCREL,"Skeleton Parsing Manual", 1989

5. An example of a piece of text demonstrating all these features using the C6 tagset and the Lancaster underscore convention for attaching word-class tags would be:-

```
[N Brezhnev_NP1 N][V said_VVD [Fn[N progress_NN1 N]@ [V was_VBDZ
needed_VVN V] @[P in_II [Tg[Tg& controlling_VVG [N the_AT
East-West_JJ [ arms_NN2 race_NN1 ]N]Tg&] ,_YCOM [Tg+ eliminating_VVG
"_YQUO [N conflict_NN1 situations_NN2 "_YQUO [P in_II [N[N& the_AT
Middle_NP1 East_NP1 N&] and_CC [N+ Southeast_NP1 Asia_NP1 N+]N]P]N]Tg+]
and_CC [Tg+ transforming_VVG [N the_AT Indian_NP1 Ocean_NP1 N][P into_II
[N a_AT1 "_YQUO [ zone_NN1 [P of_IO [N peace_NN1 N]P]] "_YQUO
[P outside_II [N great-power_JJ military_JJ rivalries_NN2
N]P]N]P]Tg+]Tg]P]Fn]V] ._YSTP
```

3.3 Selection of Text Samples

The form of analysis represented by skeleton parsing is linguistically most interesting when applied to connected text. Skeleton parsing is "sentence oriented" and so is ideally suited to selection of samples at the BNC segment level. Texts from the "core" corpus with connected text would be identified and within those a selection of appropriate segments chosen. Within such a small overall sample it would only be possible to achieve a limited degree of representativeness across different text types.

3.4 Proposed Delivery Format

1. From discussions between Lou Burnard (OUCS) and Roger Garside (Lancaster) it appears that OUCS are happy to convert enrichment annotations in Lancaster output formats (such as in the example in 3.2(5) above) to CDIF conformant equivalents after delivery.

2. The proposed delivery format is therefore as complete BNC texts in the normal Lancaster `C_file` format, but with appropriately identified skeleton parsed segments in Lancaster output format. The identification scheme could use a segment attribute, a special tag or other mechanism specified by OUCS.

   e.g.:-

   ```
   <s n=00045>
   <parsed>
   [N Brezhnev_NP1 N][V said_VVD [Fn[N progress_NN1 N]@ [V was_VBDZ
   needed_VVN V] @[P in_II [Tg[Tg& controlling_VVG [N the_AT
   East-West_JJ ...           ...military_JJ rivalries_NN2
   N]P]N]P]Tg+]Tg]P]Fn]V] ._YSTP
   </parsed>
   ```

3. The text containing the skeleton parsed sentences could either be the only delivered version of a text, or could be delivered in addition to a normal `C_file` (with C6 tagging).