BRITISH NATIONAL CORPUS

PCW44

Core Corpus Design

DRAFT

Elizabeth Eyes & Geoffrey Leech

This document is a draft to be submitted to the Project Committee meeting of 7th July 1993.

1. General Design

The Core Corpus of the British National Corpus is being compiled by the Unit for Computer Research on the English Language (UCREL) at Lancaster University. It will consist of 2M words of written and spoken text, extracted from the 100M words of the entire Main Corpus. As already agreed, the Core Corpus will be tagged using the BNC Tagset C6 (appended). The automatic tagging will be 100% manually post-edited.

Reference is made in this document to:

BNCW08 Written Corpus Design Specification
J. Clear, 2 Sept 1991

TGAW14 Spoken Corpus Design Specification
S. Crowdy et al, 24 Oct 1991

The text samples incorporated in the Core Corpus will be a subset of those in the Main Corpus. There will be no subdivision of these samples.

The target breakdown is:

50% Written
50% Spoken

In other respects, the aim is that the selection of data for the Core Corpus will correspond to the proportions of the Main Corpus, as far as can be reasonably managed, given constraints of time, ongoing availability of texts, retrieval of information from the BNC Database etc.

2. The Written Corpus

2.1 Selection Features

2.1.1 Domain

The BNC division into informative and imaginative texts will be retained in the Core Corpus:

70-80% informative
20-30% imaginative

It may not be possible to undertake sampling representative of the breakdown of the eight informative domains proposed in BNCW08 5.3.1 but the selection will be monitored to give as good a representation of these domains as possible.

2.1.2 Time

No texts will be included which were published before 1975. The dates of texts will be monitored as far as possible to give a reasonable spread over the period 1975-1993.

2.1.3 Medium

The subdivision will be:

	55-65%	books
	30-40%	periodicals
not less than	4%	plays/speeches
not less than	1%	ephemera

2.1.4 Level

Features of Level will be monitored with the aim of achieving as far as possible the same proportions as in the Main Corpus. (See BNCW08 5.3.4 Level)

2.2 Envisaged Breakdown of Written Corpus

BOOKS
Informative 400K words
Imaginative 200K words

BROCHURES
Informative 50K words

Sub-Total:
Informative 450K words
Imaginative 200K words

PERIODICALS
Informative 270K words
Imaginative 30K words

PLAYS/SPEECHES
Informative 20K words
Imaginative 20K words

EPHEMERA
Informative 10K words

Sub-total:
Informative 300K words
Imaginative 50K words

TOTAL:
Informative 750K words
Imaginative 250K words

3 The Spoken Corpus

This section of the Core Corpus will also mirror the proportions of the Main Corpus (see TGAW14 Figure 3) as far as is practicable.

Total 1M words:

Demographic samples	500K words
Context-governed samples	500K words

3.1 Demographic Samples

The aim will be to achieve representative sampling (matching that of the Main Corpus) according to:

GENDER	M/F
AGE	15-24, 24-34, 35-44, 45-60, 60+
CLASS	A/B, C1, C2, D/E
REGION	North, Midlands, South (the three super-regions)

3.2 Context-defined Samples

Breakdown (see TGAW14 Figure 3)

Educational/Informative	125K words
Business	125K words
Public/Institutional	125K words
Leisure	125K words
TOTAL	500K words

The proportions of monologue/dialogue will be monitored with the aim of achieving samples proportionate to those in the Main Corpus.

EXPLANATORY NOTE

It is clear that we are in a race against time to complete the Core Corpus, as well as the BNC Main Corpus. This means that we have had to start the C6 tagging and post-editing of the Core Corpus already. This, in turn, means that we have to continue selecting samples for the Core Corpus from now onwards, although the Main Corpus is incomplete and information is incomplete about the BNC samples which have already been collected. Some Main Corpus texts (eg Ephemera) will not be available for processing until late in the remaining time schedule for the project. Therefore these, in particular, can only be included hypothetically in the Core Corpus at present.

For all these reasons, the present design of the Core Corpus remains approximate in order to give reasonable flexibility for selection of "Core" texts in the remaining months of the

project.

EJE/GNL/28.6.93