British National Corpus

Corpus Design Specification

24 May 1991

DRAFT

1 Purpose

While it is a truism that one cannot design a corpus without having specified what purpose the corpus is to serve, in the case of the BNC it is clear that the range of possible purposes that it might serve is so wide that one is forced to fall back on the formula that it is to be a general-purpose corpus.

A great deal of the expertise and experience which the consortium can call on for the BNC has been acquired in the course of building and working with corpora for lexicography and related applications. The demands of lexicography and the compilation of language reference products will inevitably be strongly represented by the publishing houses in the consortium, who are best placed to make direct use of the corpus and thereby propagate the benefit to be derived from it into the language industries through better lexicons and dictionaries.

However, since the corpus is to be made widely available for research and development it should fulfil a wider function beyond the immediate concerns of the dictionary makers. Most of the applications which we envisage for the corpus outside the consortium are listed in the paper *Uses of the British National Corpus*. Since the consortium cannot prescribe uses for the corpus and since it cannot speak for the whole spectrum of specialists who might have a use for a large text database, we are constrained to design the corpus in such a way that it suits the primary purposes (those which the consortium identifies for itself) and is not thereby rendered unusable for other purposes which the project has not specifically identified.

2 General Definitions

The BNC will be:

- a **sample** corpus: text samples approximately 40,000 words; samples taken randomly from beginning, middle, or end; samples adjusted to convenient text breakpoint (e.g. chapter, section, paragraph).

- a **diachronic** corpus: language of the late 20th century. The corpus will include imaginative texts from 1960; informative texts after 1975.

- a **general** corpus: not specifically restricted to any particular subject field, register or genre.

- a **monolingual** British English corpus: it includes text samples which are substantially the product of speakers of British English. A marginal proportion of the words in the corpus will be in a foreign language or non-British English inasmuch as texts selected will include embedded material or material of mixed or uncertain provenance.

- a corpus of **adult** language: the written component will include writing by adults and intended for an readership at secondary school level and upwards. The spoken component can be expected to include a proportion of language produced by and addressed to children.

3 Written and Spoken Language

One of the primary decisions concerning design is the relative proportions in the BNC of written and spoken material. There is a broad consensus among the participants in the project and among corpus linguists that a general-purpose corpus of the English language would ideally contain a high proportion of spoken language relative to written texts. The ICE, for example, allocates 50% to spoken data for its corpora.[1] Since it is significantly more expensive to record and transcribe natural speech than it is to acquire written text in computer form, the spoken component of the BNC will constitute approximately 10% (10m words) of the total and the written component will be 90% (90m words). This is agreed to be a realistic target, given the constraints of time and budget, yet large enough to yield valuable empirical statistical data about spoken English. Ten million words of transcribed speech is more than is held in any current corpus, apart from highly domain-specific sources such as the Canadian Hansard data.

4 Production and Reception

While it is sometimes useful to distinguish in theory between the universe of English language as the totality of text which is *received* by people reading and listening, and that which is *produced* by people as writers and speakers, it is agreed within this project that the sampling of texts for a general-purpose corpus must take account of both perspectives and with different emphasis for speech and writing.

4.1 Written Texts

Because written texts are recorded, their reception and production are asynchronous. The storage and distribution of written texts (mainly on paper, but increasingly electronically) creates a situation in which the relationship between receiver and producer is static and the product of one or a few writers is consumed by a potentially infinite number of readers.

[1] The ICE corpora (a number of discrete 1-million word sample corpora) are designed to facilitate the comparison and contrast of international regional varieties of English.

Broadcast speech is similar to writing with respect to the relationship between producer and receiver. Television and radio, having such high cultural significance, are assumed to be very important as components in the linguistic environment and the transmission of language (both spoken and 'written to be spoken') via these media is closely related to the industry of paper and print.

In planning the sampling of texts for the BNC it became increasingly clear that very few objective measures of the scope and nature of the target population (modern, adult British English) are available upon which to base a sampling frame. There are some accessible statistics and catalogues which could be taken into account, but it was felt that none was an appropriate basis for constructing the BNC. Some of these measures are:

Books published per annum. Numbers of books published (divided into subject domains) gives some indication of activity within the publishing industry and, arguably, of written language *production*. However, it is obvious that thousands of books are published but hardly read and that authoring books is a rather specialized language activity in which very few people engage.

Books in print. This measure has a diachronic aspect, reflecting to some extent the 'durability' of some written texts. Looked at over many years, it might act as a guide to books which are perceived, at least by the publishers, to be of enduring interest.

Current magazines/periodicals. As with books, the number and variety of periodicals in circulation is no reliable guide to the production of written language since it reflects only the small proportion of writing which passes through the publishing machine to find its way into print. It may be that periodicals are bought and read by a wider cross-section of the community than books and consequently that the range of subject domains and their relative proportions may serve as an indication of the characteristics of language *reception* among speakers of British English.

Bestseller lists. This obviously focusses on language reception, and distinguishes certain texts from the mass which are published. These texts may be thought to be more influential or important linguistically, because of their wide readership. Unfortunately, the statistics of book-buying are such that a very few texts achieve very high sales while a vast number sell only a few or in modest numbers. If this distribution curve were to be adopted in the sampling frame for the BNC, then the *range* of texts which would be included would be severely reduced in order to accommodate the correct proportions for the types of text which appear in bestseller lists. If, however, a text from one particular subject domain is required for the corpus, then it may be most appropriate to select one which has achieved high sales.

Periodical circulation figures. These figures also show that a few titles dominate with very high circulation. To represent these titles in direct proportion to all others in the BNC would, we believe, result in a corpus which would not contain sufficient quantities of a range of text sources to enable meaningful contrastive analysis to be carried out.

Library lending statistics. Books most frequently borrowed or authors most frequently borrowed might be regarded as higher priority candidates for inclusion in the corpus, to reflect their apparently wider reception. As with the books in print listings, this data takes account of the 'shelf-life' of titles, though more specifically related to reception.

The available data is concerned almost exclusively with published books and periodicals. It is much more difficult to obtain data concerning the production or reception of unpublished writing. Intuitive estimates are therefore required in order to establish some guidelines for text sampling.

4.2 Spoken Texts

The situation with speech is quite different and needs separate treatment. The project plans to collect 10 million words of spoken text. For the purposes of this project we distinguish text which is written to be spoken (e.g. TV or drama scripts, scripted speeches) as part of the written component. All other text produced orally is considered in the class of spoken text, even though the degree of spontaneity of speech in this class may vary widely. Since speech is primarily ephemeral, the emphasis in collecting data is towards production. The project aims to collect at least half of the spoken component by direct recording of selected subjects over a two-day period of their normal daily activities. It is widely agreed that corpus-based study of English suffers from a bias towards the written language and that a significant proportion of the BNC should therefore be orthographically transcribed, non-elicited speech.

This approach to data collection, while seeming to fulfil requirements for a sample of authentic speech production, was felt to focus too narrowly on what might be called 'domestic' interaction and unlikely to contain any useful samples of speech from less personal, yet culturally significant, speech behaviour. TV and radio broadcasting in particular is known to have a very large audience in Britain and can be considered as a highly influential factor in the development of British English. From the point of view of the reception of the spoken language, broadcast discussion, interviews, commentary, lectures and sermons are very important components in the average British speaker's exposure to language. Other speech situations (e.g. formal meetings, telephone conversations, classroom tuition) which might be expected to occur in the normal course of people's everyday life will also be recorded to ensure that a wide variety of spoken language is captured.

5 Written Component

5.1 Sampling Methods

The approach to be adopted is one of stratified sampling. Four dimensions of variability are identified and texts are selected to fill quotas assigned to categories within each dimension. These dimensions are referred to as *selection features* and are intended to act as a control on the collection of text, to ensure that a broad range of different styles of language are captured within the corpus. Since very little empirical research has been carried out into the distribution of linguistic features over large samples of authentic text, it is impossible at this stage to define the extent to which different types of text (or even different structural units within a text) will show differing linguistic characteristics. Without any firm guides as to how the sampling strategy will affect the results of linguistic analyses, our approach is to collect at least some examples of almost all types of text. We intend to enable two types of analysis.

The first takes the corpus to be a microcosm of modern British English and derives linguistic information which can be taken, with some degree of confidence, to exemplify real patterns

of usage. For this purpose it is important to gather texts from a wide variety of sources in order that any statistics obtained can be held to be valid for more than merely one or a few specific text sources.

The second type of analysis is contrastive and involves comparing different components of the corpus to investigate differences and similarities between types of text. For this purpose too it is valuable to have a wide variety of texts, sufficient to enable reasonably large[2] sub-corpora to be defined and compared with other configurations.

The four selection features are divided into classes, and texts are selected to fill the target percentages detailed below. The target percentages were arrived at by shared intuitive estimate of the nature of general modern British English and are not directly drawn from any objective measure of the composition of the vast and ill-defined population to be sampled. They are expressed as a range, since it is clear that there is no good justification for precise figures. The range represents the maxima and minima which the project considers to be appropriate for the range of purposes envisaged for the corpus.

In order to provide some sort of 'control group' against which to evaluate the intuitive stratification of the sampling, it is agreed that one half of the texts collected in the Books and Periodicals class of the medium feature should be selected randomly from comprehensive catalogues (e.g. *Books in Print*, *Willings Press Guide*) without regard to the target percentages for Domain, Time or Level features. It is not considered feasible to carry out this kind of random selection for unpublished writing. Once a text is randomly picked, it will be classified according to the selection features. As a consequence of collecting some of the material by this method, the overall percentages of text in each of the selection categories may be altered according as the randomly collected material differs in its composition from the targeted selection.

5.2 Selection Features

- Domain
- Time
- Medium
- Level

5.2.1 Domain

The classification of texts according to subject fields seems hardly appropriate to texts which are fictional or which are generally perceived to be literary or creative. Consequently, these texts are distinguished under the label IMAGINATIVE and are not assigned to particular subject areas. All other texts are treated as INFORMATIVE and are assigned to one of the 9 domain headings listed below. The evidence from catalogues of books and periodicals suggests that imaginative texts account for significantly less than 25% of published output and unpublished reports, correspondence, reference works and so on would seem to add further to the bulk of informative text which is produced and consumed. However, the overall distribution between informative and imaginative text samples is set to reflect the influential cultural role of literature and creative writing. The target percentages for the

[2] in the order of, say, 1 million words

nine informative domains approximate closely to the pattern of book publishing in the UK during the past 20 years or so, as manifest in the categorised figures for new publications that appear annually in Whitaker's Book List.

	%age Written	Overall (words)
INFORMATIVE	60–80%	54–72m
IMAGINATIVE	20–40%	18–36m

INFORMATIVE

	Domain	%age of Written	(words)
1	Natural & pure science	5%	4.5m
2	Applied Science	5%	4.5m
3	Social science	15%	13.5m
4	World affairs	12%	10.8m
5	Commerce & finance	10%	9m
6	Arts (rock & pop, dance, theatre,...)	8%	7.2m
7	Belief & thought (religion, philosophy,...)	4%	3.6m
8	Leisure (sport, gardening,...)	10%	9m
9	Biography	5%	4.5m

5.2.2 Time

For published material, the date of first publication is taken, even though it is possible for texts to have been written a long time before publication. For unpublished texts, the best available evidence is taken as to the date of writing. The relative proportions between the two date categories are different for informative and imaginative text samples, reflecting the longer 'shelf-life' of imaginative writing.

	Imaginative (20–40%)	Informative (60–80%)	Overall (words)
1960–1974	25%	-	4.5–9m
1975–	75%	100%	67.5–81m

5.2.3 Medium

This categorisation is broad, since a detailed taxonomy or feature classification of text medium could involve the proliferation of subcategories to the extent that it becomes impossible for the BNC adequately to represent each detailed category. The labels here are intended to be comprehensive in the sense that any text can be assigned with reasonable confidence to these macro categories. The labels we have adopted represent the highest levels of a fuller taxonomy of text medium.

Books, Periodicals*
- Books 50–70%*
- Periodicals 20–30%*

Miscellaneous

	Published	5–10%
		brochures, leaflets, manuals, adverts
	Unpublished	5–10%
		letters/memos, reports, minutes, essays
To be spoken		2–7%

The component which is Books and Periodicals will be made up of two segments, one collected by random selection from relevant catalogues of publications and the other by manual selection according to the selection features.

5.2.4 Level

The Level feature has a slightly different meaning in relation to the Imaginative and Informative categories. For imaginative texts the level indicates how a text is regarded with respect to its literary and intellectual character. For informative texts the levels denote the extent to which the text is aimed at a technical and specialised readership, the informed layperson or a very wide popular audience.

These categories are intuitively strongly felt, but difficult to define objectively. The three-level categorisation is agreed to be workable and useful. The percentages given below reflect the view that imaginative texts (novels, drama, stories, etc.) are more susceptible to variation in level, whereas informative texts at the low level will be more rare. Moreover, since the BNC is aiming to sample a general level of British English usage, the precentage for the high technical level has been reduced in favour of texts for the layperson.

	Imaginative	Informative	Overall
High	33%	30%	31% (28m words)
Middle	33%	50%	45% (40m words)
Low	33%	20%	23% (21m words)

Notes

These percentages may vary by 5% up or down.

The Overall figures assume a distribution of 25% Imaginative, 75% Informative.

5.3 Classification Features

In addition to the controlled text sampling according to selection features, texts are classified with respect to a second set of *classification features*, for which no target proportions are set. This information is recorded to allow more delicate contrastive analysis of particular sets of texts. As a simple example, the gross division into two time periods in the selection features can, of course, be refined and subcorpora defined over the BNC for more specific dates. However, the relative sizes of such subcorpora are undefined by the BNC design specification.

The classification features are:

- sample id

- sample size (words)
- sample extent (start/end points)
- text size (words)
- text composition (single, composite, collection)
- standard bibliographic reference
- data capture history
- subject field
- authorship (multiple, corporate, unknown)
- author gender
- author age group
- author ethnic group
- author domicile
- target age group
- ?target gender

6 Spoken Component

The project plans to collect a spoken component of 10 million words. It should be pointed out that this is in no way to be equated with a *speech* corpus: this corpus material is to be recorded at quality levels that will permit accurate orthographic transcription, but is not intended to fulfil the requirements of speech researchers for very high quality recording for acoustic or phonetic data. The transcription will be carried out by keyboarders at the orthographic level, since we hope to capture the lexical and grammatical characteristics of spoken language rather than prosodic or phonetic.

The audio tapes which are produced will be archived and can be made available to organisations outside the project. There are no plans to manipulate the audio recordings in any way beyond what is required to carry out the transcription effectively.

6.1 Random Sampling

At least 5 million words of speech is to be collected by an approach which would involve the identification of 100–200 demographically selected subjects, native speakers of British English from various regional, age and educational categories. These subjects would carry portable sound-recording equipment around with them for two 24-hour periods, recording all those with whom they came into contact during the time that recording is in progress. The subjects would obviously be fully aware that the recording was being made, but their interlocutors would not know until after the event—in order to capture spontaneous and natural speech.

Several small-scale pilots have been carried out, with perfectly acceptable recording quality and accurate transcriptions being achieved. Exhaustive proofreading will not be carried out, since this would be too costly and time-consuming. Automatic spell-checking and manual spot checks will be carried out after transcription.

The recordings will be transferred to Digital Audio Tape (DAT) and edited to remove long silences and inaudible segments.

6.2 Stratified Sampling

In order to supplement the collection of primarily private spontaneous speech, at least 2 million words of material will be collected to fill the text type classification outlined below.

- Dialogue
  - Private
    - Face-to-face conversation
      - Structured
      - Unstructured
    - Distanced conversation
    - Classroom interaction
  - Public
    - Broadcast discussion/debate
    - Legal proceedings
- Monologue
  - Commentary
  - Unscripted speeches
  - Demonstrations

As with the collection of written material, it is expected that the material gathered by random sampling can potentially cut across all varieties of speech situation, but the relative proportions cannot be predicted to any useful extent. It is unlikely that any of the selected subjects will give a formal speech or a television interview during the two-day recording period, so this type of text, which should not be completely absent from the BNC, is to be collected by deliberate means.

6.3 Classification Features

These are (provisionally):

Reference Information

- reference number
- time/date
- location
- recording information

Details of Speech Event

- setting (workplace, train, home)
- discourse type (conversation, interview, lecture, telephone)
- spontaneity factor (partially scripted, unscripted preplanned, spontaneous)

Details of Participants

- name/identifier
- sex
- age
- ethnic origin
- region/accent
- occupation
- education
- social class
- relationship to interlocutor(s) (family, intimate friend, colleague, etc)
- status relative to interlocutor (high/low)
- other details (speech impairment, other situational factors)

Domain of Speech Event