Corpus Linguistics Basics

Crash course for SFB-991 members

22.01.2013

Overview

- What sort of a corpus is the BNC
- Design of the BNC: composition
- Representativeness, balance and sampling revisited
- Design of the BNC written component: sampling
- Design of the BNC written component: descriptive features
- Design of the BNC written component: selection procedure
- Case study: forensic linguistics

Overview

- What sort of a corpus is the BNC
- Design of the BNC: composition
- Representativeness, balance and sampling revisited
- Design of the BNC written component: sampling
- Design of the BNC written component: descriptive features
- Design of the BNC written component: selection procedure
- Case study: forensic linguistics

WHAT SORT OF CORPUS IS THE BNC?

- Monolingual: It deals with modern British English, not other languages used in Britain. However non-British English and foreign language words do occur in the corpus.
- Synchronic: It covers British English of the late twentieth century, rather than the historical development which produced it.
- General: It includes many different styles and varieties, and is not limited to any particular subject field, genre or register. In particular, it contains examples of both spoken and written language.
- Sample: For written sources, samples of 45,000 words are taken from various parts of single- author texts. Shorter texts up to a maximum of 45,000 words, or multi-author texts such as magazines and newspapers, are included in full. Sampling allows for a wider coverage of texts within the 100 million limit, and avoids over-representing idiosyncratic texts.

BNC-Guide

Overview

- What sort of a corpus is the BNC
- **Design of the BNC: composition**
- Representativeness, balance and sampling revisited
- Design of the BNC written component: sampling
- Design of the BNC written component: descriptive features
- Design of the BNC written component: selection procedure
- Case study: forensic linguistics

BNC User Reference Guide: 1 Design of the corpus. Composition.

http://www.natcorp.ox.ac.uk/docs/URG/BNCdes.html

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 in relation to written texts.

However, it is significantly more expensive to record and transcribe natural speech than to acquire written text in computer-readable form.

Consequently the spoken component of the BNC constitutes approximately 10 per cent (10 million words) of the total and the written component 90 per cent (90 million words).

These were agreed to be realistic targets, given the constraints of time and budget, yet large enough to yield valuable empirical statistical data about spoken English.

In the BNC sampler, a two per cent sample taken from the whole of the BNC, spoken and written language are present in approximately equal proportions, but other criteria are not equally balanced.

BNC User Reference Guide: 1 Design of the corpus. Composition.

http://www.natcorp.ox.ac.uk/docs/URG/BNCdes.html

From the start, a decision was taken to select material for inclusion in the corpus according to an overt methodology, with specific target quantities of clearly defined types of language.

This approach makes it possible for other researchers and corpus compilers to review, emulate or adapt concrete design goals.

This section outlines these design considerations, and reports on the final make-up of the BNC. This and the other tables in this section show the actual make-up of the second version of the British National Corpus (the BNC World Edition) in terms of

- texts : number of distinct samples not exceeding 45,000 words
- S-units: number of <s> elements identified by the CLAWS system (more or less equivalent to sentences)
- W-units: number of <w> elements identified by the CLAWS system (more or less equivalent to words)

BNC User Reference Guide: 1 Design of the corpus. Composition.

http://www.natcorp.ox.ac.uk/docs/URG/BNCdes.html

The XML Edition of the BNC contains 4049 texts and occupies (including all markup) 5,228,040 Kb, or about 5.2 Gb.

In total, it comprises just under 100 million orthographic words (specifically, 96986707), but the number of w-units (POS-tagged items) is slightly higher at 98363783.

The tagging distinguishes a further 13614425 punctuation strings, giving a total content count of 110691482 strings.

The total number of s-units tagged is about 6 million (6026284). Counts for these and all the other elements tagged in the corpus are provided in the corpus header.

In the following tables both an absolute count and a percentage are given for all the counts. The percentage is calculated with reference to the relevant portion of the corpus, for example, in the table for "written text domain", with reference to the total number of w-units in written texts. Note that punctuation strings are not included in these totals. The reference totals used are given in the first table below.

BNC User Reference Guide: 1 Design of the corpus. Composition.

http://www.natcorp.ox.ac.uk/docs/URG/BNCdes.html

Table 1. Text type

	texts	w-units	%	s-units	%
spoken demographic	153	4233955	4.30	610557	10.13
spoken context-governed	755	6175896	6.27	427523	7.09
written books and periodicals	2685	79238146	80.55	4395581	72.94
written-to-be-spoken	35	1278618	1.29	104665	1.73
written miscellaneous	421	7437168	7.56	487958	8.09

BNC User Reference Guide: 1 Design of the corpus. Composition.

http://www.natcorp.ox.ac.uk/docs/URG/BNCdes.html

Table 2. Publication date

	texts	w-units	%	s-units	%
Unknown	162	183158	1.86	126416	2.09
1960-1974	46	1718449	1.74	119510	1.98
1975-1984	169	4730889	4.80	257962	4.28
1985-1993	3672	90082860	91.58	5522396	91.63

Overview

- What sort of a corpus is the BNC
- Design of the BNC: composition
- Representativeness, balance and sampling revisited
- Design of the BNC written component: sampling
- Design of the BNC written component: descriptive features
- Design of the BNC written component: selection procedure
- Case study: forensic linguistics

Representativeness, Balance and Sampling Revisited

Representativeness refers to the extent to which a sample includes the full range of variability in a language and language variety.

The representativeness of a (general) corpus depends on two factors:

- balance or the range of genres and registers included in the corpus;
- sampling techniques or how the text excerpts for each genre are selected.

Representativeness, balance and sampling revisited

Some aspects of the representativeness:

- The criteria used to select the texts for a certain corpus have to be external (non-linguistic). One of the main uses of corpora is to examine naturally occurring linguistic feature distributions. The results of corpus analyses can be used to improve its representativeness and to discover design lapses and errors.
- Broad range of genres is essential for general corpora.
- Production and reception are important aspects of language usage and have to be balanced in a corpus.

Representativeness, balance and sampling revisited

Sampling:

- sampling unit, e.g. a book, periodical or newspaper;
- sampling frame – the list of sampling units, e.g., catalogues or bibliographies;
- sampling techniques, e.g., simple random sampling, stratified random sampling (proportionality is an issue by stratified sampling);
- sample size – full text vs. text chunks.

Representativeness, Balance and Sampling Revisited

In order to obtain a representative sample from a population, the first concern to be addressed is to define the *sampling unit* and the boundaries of the population. For written text, for example, a sampling unit may be a book, periodical or newspaper. The population is the assembly of all sampling units while the list of sampling units is referred to as a *sampling frame*.

In corpus design, a population can be defined in terms of language production, language reception or language as a product. The first two designs are basically demographically oriented as they use the demographic distribution (e.g, age, sex, social class) of the individuals who produce/ receive language data to define the population while the last is organized around text category/genre of language data.

However, it can be notoriously difficult to define a population or construct a sampling frame, particularly for spoken language, for which there are no ready-made sampling frames in the form of catalogues or bibliographies.

Corpus-Based Language Studies, 19-21

Representativeness, Balance and Sampling Revisited

Once the target population and the sampling frame are defined, different sampling techniques can be applied to choose a sample which is as representative as possible' of the population.

A basic sampling method is simple random sampling. [Another method is] stratified random sampling.

A further decision to be made in sampling relates to sample size. For example, with written language, should we sample full texts (i.e. whole document s) or text chunks? If text chunks are to be sampled, should we sample text initial, middle or end chunks?

Another sampling issue, which particularly relates to stratified sampling, is the proportion and number of samples for each text category. The numbers of samples across text categories should be proportional to their frequencies and/or weights in the target population in order for the resulting corpus to be considered as representative.

Corpus-Based Language Studies, 19-21

Overview

- What sort of a corpus is the BNC
- Design of the BNC: composition
- Representativeness, balance and sampling revisited
- Design of the BNC written component: sampling
- Design of the BNC written component: descriptive features
- Design of the BNC written component: selection procedure
- Case study: forensic linguistics

BNC User Reference Guide: 1.4 Design of the written component. Sampling. http://www.natcorp.ox.ac.uk/docs/URG/BNCdes.html#wrides

Sampling basis: production and reception

While it is sometimes useful to distinguish in theory between language which is received (read and heard) and that which is produced (written and spoken), it was agreed that the selection of samples for a general- purpose corpus must take account of both perspectives.

Text that is published in the form of books, magazines, etc., is not representative of the totality of written language that is produced, as writing for publication is a comparatively specialized activity in which few people engage.

BNC User Reference Guide: 1.4 Design of the written component. Sampling. http://www.natcorp.ox.ac.uk/docs/URG/BNCdes.html#wrides

However, it is much more representative of written language that is received, and is also easier to obtain in useful quantities, and thus forms the greater part of the written component of the corpus.

There was no single source of information about published material that could provide a satisfactory basis for a sampling frame, but a combination of various sources furnished useful information about the totality of written text produced and, particularly, received, some sources being more significant than others.

They are principally statistics about books and periodicals that are published, bought or borrowed.

BNC User Reference Guide: 1.4 Design of the written component. Sampling. http://www.natcorp.ox.ac.uk/docs/URG/BNCdes.html#wrides

Catalogues of books published per annum tell us something about production but little about reception as many books are published but hardly read.

A list of books in print provides somewhat more information about reception as time will weed out the books that nobody bought (or read): such a list will contain a higher proportion of books that have continued to find a readership.

The books that have the widest reception are presumably those that figure in bestseller lists, particularly prize winners of competitions such as the Booker or Whitbread.

BNC User Reference Guide: 1.4 Design of the written component. Sampling. http://www.natcorp.ox.ac.uk/docs/URG/BNCdes.html#wrides

Such works were certainly candidates for inclusion in the corpus, but the statistics of book-buying are such that very few texts achieve high sales while a vast number sell only a few or in modest numbers.

If texts had been selected in strict arithmetical proportion to their sales, their range would have been severely limited. However, where a text from one particular subject domain was required, it was appropriate to prefer a book which had achieved high sales to one which had not.

Library lending statistics, where these are available, also indicate which books enjoy a wide reception and, like lists of books in print, show which books continue to be read.

BNC User Reference Guide: 1.4 Design of the written component. Sampling. http://www.natcorp.ox.ac.uk/docs/URG/BNCdes.html#wrides

Similar observations hold for magazines and periodicals. lists of current magazines and periodicals are similar to catalogues of published books, but perhaps more informative about language reception, as it may be that periodicals are bought and read by a wider cross-section of the community than books. Also, a periodical that fails to find a readership will not continue to be published for long.

Periodical circulation figures have to be treated with the same caution as bestseller lists, as a few titles dominate the market with a very high circulation. To concentrate too exclusively on these would reduce the range of text types in the corpus and make contrastive analysis difficult.

BNC User Reference Guide: 1.4 Design of the written component. Sampling.

http://www.natcorp.ox.ac.uk/docs/URG/BNCdes.html#wrides

Published written texts were selected partly at random from Whitaker's Books in Print for 1992 and partly systematically, according to the selection features outlined in section Selection features below.

Available sources are 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 were therefore made in order to establish some guidelines for text sampling in the latter area.

BNC User Reference Guide: 1.4 Design of the written component. Sampling.

http://www.natcorp.ox.ac.uk/docs/URG/BNCdes.html#wrides

Selection features

Texts were chosen for inclusion according to three selection features:

- domain (subject field),
- time (within certain dates) and
- medium (book, periodical, etc.).

The purpose of these selection features was to ensure that the corpus contained a broad range of different language styles, for two reasons. The first was so that the corpus could be regarded as a microcosm of current British English in its entirety, not just of particular types. The second was so that different types of text could be compared and contrasted with each other.

BNC User Reference Guide: 1.4 Design of the written component. Sampling.

http://www.natcorp.ox.ac.uk/docs/URG/BNCdes.html#wrides

Selection Procedure

Each selection feature was divided into classes (e.g. 'Medium' into books, periodicals, unpublished etc.; 'Domain' into imaginative, informative, etc.) and target percentages were set for each class.

These percentages are quite independent of each other: there was no attempt, for example, to make 25 per cent of the selected periodicals imaginative.

The design proposed that seventy-five per cent of the samples be drawn from informative texts, and the remaining 25 per cent from imaginative texts.

BNC User Reference Guide: 1.4 Design of the written component. Sampling. http://www.natcorp.ox.ac.uk/docs/URG/BNCdes.html#wrides

It further proposed that titles be taken from a variety of media, in the following proportions:

- 60 per cent from books,
- 30 per cent from periodicals,
- 10 per cent from miscellaneous sources (published, unpublished, and written to be spoken).

Half of the books in the 'Books and Periodicals' class were selected at random from Whitaker's Books in Print 1992. This was to provide a control group to validate the categories used in the other method of selection: the random selection disregarded Domain and Time, but texts selected by this method were classified according to these other features after selection.

BNC User Reference Guide: 1.4 Design of the written component. Sampling. http://www.natcorp.ox.ac.uk/docs/URG/BNCdes.html#wrides

Sample size and method

For books, a target sample size of 40,000 words was chosen.

No extract included in the corpus exceeds 45,000 words.

For the most part, texts which in their entirety were shorter than 40,000 words were further reduced by ten per cent for copyright reasons; a few texts longer than the target size were however included in their entirety.

Text samples normally consist of a continuous stretch of discourse from within the whole. A convenient breakpoint (e.g. the end of a section or chapter) was chosen as far as possible to begin and end the sample so that high-level discourse units were not fragmented.

BNC User Reference Guide: 1.4 Design of the written component. Sampling. http://www.natcorp.ox.ac.uk/docs/URG/BNCdes.html#wrides

Where possible, no more than one sample was taken from any one text; for newspaper texts and large encyclopaedic works, no sample greater than 40,000 words was taken.

Samples were taken randomly from the beginning, middle or end of longer texts. (In cases where a publication included essays or articles by a variety of authors of different nationalities, the work of non-UK authors was omitted.)

Some types of written material are composite in structure: that is, the physical object in written form is composed of more than one text unit. Important examples are issues of a newspaper or magazine which, though editorially shaped as a document, contain discrete texts, each with its specific authorship, stylistic characteristics, register and domain.

BNC User Reference Guide: 1.4 Design of the written component. Sampling.

http://www.natcorp.ox.ac.uk/docs/URG/BNCdes.html#wrides

The BNC attempts to separate these discrete texts where appropriate and to classify them individually according to the selection and classification features. As far as possible, the individual stories in one issue of a newspaper were grouped according to domain, for example as ‘Business’ articles, ‘Leisure’ articles, etc.

The following subsections discuss each selection criterion, and indicate the actual numbers of words in each category included.

BNC User Reference Guide: 1.4 Design of the written component. Sampling.

http://www.natcorp.ox.ac.uk/docs/URG/BNCdes.html#wrides

Domain

Classification according to subject field seems hardly appropriate to texts which are fictional or which are generally perceived to be literary or creative.

Consequently, these texts are all labelled imaginative and are not assigned to particular subject areas.

All other texts are treated as informative and are assigned to one of the eight domains listed below.

BNC User Reference Guide: 1.4 Design of the written component. Sampling.

http://www.natcorp.ox.ac.uk/docs/URG/BNCdes.html#wrides

Table 3. Written Domain

	texts	w-units	%	s-units	%
Imaginative	476	16496420	18.75	1352150	27.10
Informative:natural & pure science	146	3821902	4.34	183384	3.67
Informative: applied science	370	7174152	8.15	356662	7.15
Informative: social science	526	14025537	15.94	698218	13.99
Informative: world affairs	483	17244534	19.60	798503	16.00
Informative: commerce & finance	295	7341163	8.34	382374	7.66
Informative: arts	261	6574857	7.47	321140	6.43
Informative: belief & thought	146	3037533	3.45	151283	3.03
Informative: leisure	438	12237834	13.91	744490	14.92

BNC User Reference Guide: 1.4 Design of the written component. Sampling. http://www.natcorp.ox.ac.uk/docs/URG/BNCdes.html#wrides

The evidence from catalogues of books and periodicals suggests that imaginative texts account for significantly less than 25 per cent 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 eight informative domains were arrived at by consensus within the project, based loosely upon the pattern of book publishing in the UK during the past 20 years or so, as reflected in the categorized figures for new publications that appear annually in Whitaker's Book list.

Overview

- What sort of a corpus is the BNC
- Design of the BNC: composition
- Representativeness, balance and sampling revisited
- Design of the BNC written component: sampling
- Design of the BNC written component: descriptive features
- Design of the BNC written component: selection procedure
- Case study: forensic linguistics

BNC User Reference Guide: 1.4 Design of the written component. Descriptive features.

http://www.natcorp.ox.ac.uk/docs/URG/BNCdes.html#wrides

Descriptive features

Written texts may be further classified according to sets of descriptive features.

These features describe the sample texts; they did not determine their selection.

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.

BNC User Reference Guide: 1.4 Design of the written component. Descriptive features.

http://www.natcorp.ox.ac.uk/docs/URG/BNCdes.html#wrides

However, the relative sizes of such subcorpora are undefined by the BNC design specification.

These descriptive features were monitored during the course of the data gathering, and text selection, in cases where a free choice of texts was available, took account of the relative balance of these features.

Thus although no relative proportions were defined for different target age groups (for example), we ensured that the corpus does contain texts intended for children as well as for adults.

The following tables summarize the results for the first release of the corpus. Note that many texts remain unclassified.

BNC User Reference Guide: 1.4 Design of the written component. Descriptive features.

http://www.natcorp.ox.ac.uk/docs/URG/BNCdes.html#wrides

Author information

Information about authors of written texts was included only where it was readily available, for example from the dust-wrapper of a book. Consequently, the coverage of such information is very patchy.

The authorship of a written text was characterized as ‘corporate’ where it was produced by an organization and no specific author was given, and as ‘multiple’ in cases where several authors were named.

Author sex was classified as ‘mixed’ where more than one author of either sex was specified, and ‘unknown’ where it could not reliably be determined from the author's name.

Note that ‘author age’ means the author's age at the time of creation of the work concerned.

BNC User Reference Guide: 1.4 Design of the written component. Descriptive features.

http://www.natcorp.ox.ac.uk/docs/URG/BNCdes.html#wrides

Table 3. Author type

	texts	w-units	%	s-units	%
Unknown	211	3786835	4.30	174371	3.49
Corporate author	347	6497144	7.38	455649	9.13
Multiple author	1322	34563219	39.29	1810901	36.30
Sole author	1261	43106734	49.01	2547283	51.06

BNC User Reference Guide: 1.4 Design of the written component. Descriptive features.

http://www.natcorp.ox.ac.uk/docs/URG/BNCdes.html#wrides

The following descriptive feature are represented in similar tables:

- sex of the author;
- author age group;
- author domicile;
- target audience age;
- target audience sex;
- publication place;
- sampling type (cf. next slide).

For details, cf. :

http://www.natcorp.ox.ac.uk/docs/URG/BNCdes.html#body.1_div.1_div.4_div.3

BNC User Reference Guide: 1.4 Design of the written component. Descriptive features.
http://www.natcorp.ox.ac.uk/docs/URG/BNCdes.html#wrides

Table 12. Sampling type

	texts	w-units	%	s-units	%
Unknown	1583	35551102	40.42	1991798	39.93
Whole text	270	6524975	7.41	433722	8.69
Beginning sample	584	21075222	23.96	1119251	22.43
Middle sample	510	18454807	20.98	1049692	21.04
End sample	119	4317326	4.90	253322	5.07
Composite sample	75	2030500	2.30	140419	2.81

Overview

- What sort of a corpus is the BNC
- Design of the BNC: composition
- Representativeness, balance and sampling revisited
- Design of the BNC written component: sampling
- Design of the BNC written component: descriptive features
- Design of the BNC written component: selection procedure
- Case study: forensic linguistics

BNC User Reference Guide: 1.4 Design of the written component. Selection procedure.

http://www.natcorp.ox.ac.uk/docs/URG/BNCdes.html#wrides

Selection procedures employed

Books

Roughly half the titles were randomly selected from available candidates identified in Whitaker's Books in Print (BIP), 1992, by students of Library and Information Studies at Leeds City University. Each text randomly chosen was accepted only if it fulfilled certain criteria: it had to be published by a British publisher, contain sufficient pages of text to make its incorporation worthwhile, consist mainly of written text, fall within the designated time limits, and cost less than a set price. The students noted the ISBN, author, title and price of each book thus selected; the final selection weeded out texts by non-UK authors.

Half of the books having been selected by this method, the remaining half were selected systematically to make up the target percentages in each category. The selection proceeded as follows.

BNC User Reference Guide: 1.4 Design of the written component. Selection procedure.

http://www.natcorp.ox.ac.uk/docs/URG/BNCdes.html#wrides

Bestsellers

Because of their wide reception, bestsellers were obvious candidates for selection. The lists used were those that appeared in the Bookseller at the end of the years 1987 to 1993 inclusive.

Some of the books in the lists were rejected, for a variety of reasons. Obviously books that had already been selected by the random method were excluded, as were those by non-UK authors.

In addition, a limit of 120,000 words from any one author was imposed, and books belonging to a domain or category whose quota had already been reached were not selected.

Other bestseller lists were obtained from The Guardian, the British Council, and from Blackwells Paperback Shop.

The titles yielded by this search were mostly in the Imaginative category.

BNC User Reference Guide: 1.4 Design of the written component. Selection procedure.

http://www.natcorp.ox.ac.uk/docs/URG/BNCdes.html#wrides

Literary prizes

The criteria for inclusion were the same as for bestsellers.

The prize winners, together with runners-up and shortlisted titles, were taken from several sources, principally Anne Strachan, Prizewinning literature: UK literary award winners, London, 1989.

For 1990 onwards the sources used were: the last issue of the Bookseller for each year; The Guardian Index, 1989–, entries under the term 'Literature'; and The Times Index, 1989-, entries under the term 'Literature — Awards'.

Literary prizes are in the main awarded to works that fall into the Imaginative category, but there are some Informative ones also.

BNC User Reference Guide: 1.4 Design of the written component. Selection procedure.

http://www.natcorp.ox.ac.uk/docs/URG/BNCdes.html#wrides

Library loans

The source of statistics in this category was the record of loans under Public Lending Right, kindly provided by Dr J. Parker, the Registrar. The information comprised lists of the hundred most issued books and the hundred most issued children's books, in both cases for the years 1987 to 1993.

The lists consist almost exclusively of imaginative literature, and many titles found there also appear in the lists of bestsellers and prize winners.

BNC User Reference Guide: 1.4 Design of the written component. Selection procedure.

http://www.natcorp.ox.ac.uk/docs/URG/BNCdes.html#wrides

Additional texts

As collection proceeded, monitoring disclosed potential shortfalls in certain domains. A further selection was therefore made, based on the 'Short Loan' collections of seven University libraries. (Short Loan collections typically contain books required for academic courses, which are consequently in heavy demand.)

BNC User Reference Guide: 1.4 Design of the written component. Selection procedure.

http://www.natcorp.ox.ac.uk/docs/URG/BNCdes.html#wrides

Periodicals and magazines

Periodicals, magazines and newspapers account for 30 per cent of the total text in the corpus.

Of these, about 250 titles were issues of newspapers. These were selected to cover as wide a spectrum of interests and language as possible. Newspapers were selected to represent as wide a geographic spread as possible: The Scotsman and the Belfast Telegraph are both represented, for example.

BNC User Reference Guide: 1.4 Design of the written component. Selection procedure. http://www.natcorp.ox.ac.uk/docs/URG/BNCdes.html#wrides

Other media

In addition to samples from books, periodicals, and magazines, the written part of the corpus contains about seven million words classified as 'Miscellaneous Published', 'Miscellaneous Unpublished', or as 'Written to be spoken'.

The distinction between 'published' and 'unpublished' is not an easy one; the former category largely contains publicity leaflets, brochures, fact sheets, and similar items, while the latter has a substantial proportion of school and university essays, unpublished creative writing or letters, and internal company memoranda.

The 'written to be spoken' material includes scripted material, intended to be read aloud such as television news broadcasts; transcripts of more informal broadcast materials such as discussions or phone-ins are included in the spoken part of the corpus.

Overview

- What sort of a corpus is the BNC
- Design of the BNC: composition
- Representativeness, balance and sampling revisited
- Design of the BNC written component: sampling
- Design of the BNC written component: descriptive features
- Design of the BNC written component: selection procedure
- Case study: forensic linguistics

Derek Bentley Case

Excerpt from "… and then … Language Description and Author Attribution" by Malcolm Coulthard

In November 1952 two teenagers, Derek Bentley aged 19 and Chris Craig aged 16, were seen climbing up onto the roof of a London warehouse.

The police surrounded the building and three unarmed officers went up onto the roof to arrest them.

Bentley immediately surrendered;

Craig started shooting, wounding one policeman and killing a second.

Bentley was jointly charged with his murder, even though he had been under arrest for some time when the officer was killed.

The trial, which lasted only two days, took place five weeks later and both were found guilty. Craig, because he was legally a minor, was sentenced to life imprisonment;

Bentley was sentenced to death and executed shortly afterwards.

Bentley's family fought tenaciously to overturn the guilty verdict and were eventually successful 46 years later, in the summer of 1998.

Derek Bentley Case

Excerpt from "… and then … Language Description and Author Attribution" by Malcolm Coulthard

The evidence which was the basis for both Bentley's conviction and the subsequent successful appeal was in large part linguistic.

In the original trial the problem for the Prosecution, in making the case against Bentley, was to demonstrate that he could indeed be guilty of murder despite being under arrest when the murder was committed.

At this point it would be useful to read the statement which, it was claimed, Bentley dictated shortly after his arrest.

It is presented in full below; the only changes I have introduced are the numbering of sentences for ease of reference and the highlighting, by underlining and bold, of items to which I will later refer.

Derek Bentley Case

Excerpt from "… and then … Language Description and Author Attribution" by Malcolm Coulthard

Derek Bentley's Statement

(1) I have known Craig since I went to school. (2) We were stopped by our parents going out together, but we still continued going out with each other - I mean **we have not gone out** together until tonight. (3) I was watching television tonight (2 November 1952) and between 8 p.m. and 9 p.m. Craig called for me. (4) My mother answered the door and I heard her say that I was out. (5) I had been out earlier to the pictures and got home just after 7 p.m. (6) A little later Norman Parsley and Frank Fasey called. (7) **I did not answer the door or speak to them**.

Derek Bentley Case

Excerpt from "… and then … Language Description and Author Attribution" by Malcolm Coulthard

(8) My mother told me that they had called and I then ran out after them. (9) I walked up the road with them to the paper shop where I saw Craig standing. (10) We all talked together and then Norman Parsley and Frank Fazey left. (11) Chris Craig and I then caught a bus to Croydon. (12) We got off at West Croydon and then walked down the road where the toilets are - I think it is Tamworth Road. (13) When we came to the place where you found me, Chris looked in the window. (14) There was a little iron gate at the side. (15) Chris then jumped over and I followed. (16) Chris then climbed up the drainpipe to the roof and I followed. (17) Up to then **Chris had not said anything.** (18) We both got out on to the flat roof at the top. (19) Then someone in a garden on the opposite side shone a torch up towards us. (20) Chris said: 'It's a copper, hide behind here.' (21) We hid behind a shelter arrangement on the roof. (22) We were there waiting for about ten minutes. (23) **I did not know** he was going to use the gun. (24) A plain

Derek Bentley Case

Excerpt from "... and then ... Language Description and Author Attribution" by Malcolm Coulthard

clothes man climbed up the drainpipe and on to the roof. (25) The man said: 'I am a police officer - the place is surrounded.' (26) He caught hold of me and as we walked away Chris fired. (27) **There was nobody else** there at the time. (28) The policeman and I then went round a corner by a door. (29) A little later the door opened and a policeman in uniform came out. (30) Chris fired again then and this policeman fell down. (31) I could see that he was hurt as a lot of blood came from his forehead just above his nose.

Derek Bentley Case

Excerpt from "… and then … Language Description and Author Attribution" by Malcolm Coulthard

(32) The policeman dragged him round the corner behind the brickwork entrance to the door. (33) I remember I shouted something but I forgot what it was. (34**) I could not see** Chris when I shouted to him - he was behind a wall. (35) I heard some more policemen behind the door and the policeman with me said: **'I don't think** he has many more bullets left.' (36) Chris shouted 'Oh yes I have' and he fired again. (37) I think I heard him fire three times

Derek Bentley Case

Excerpt from "… and then … Language Description and Author Attribution" by Malcolm Coulthard

altogether. (38) The policeman then pushed me down the stairs and **I did not see** any more. (39) I knew we were going to break into the place. (40) **I did not know** what we were going to get - just anything that was going. (41**) I did not have** a gun and **I did not know** Chris had one until he shot. (42) I now know that the policeman in uniform that was shot is dead. (43) I should have mentioned that after the plain clothes policeman got up the drainpipe and arrested me, another policeman in uniform followed and I heard someone call him 'Mac'. (44) He was with us when the other policeman was killed.

(End of Bentley's statement)

Derek Bentley Case

Excerpt from “… and then … Language Description and Author Attribution” by Malcolm Coulthard

Bentley’s barrister spelled out for the jury the two necessary pre-conditions for them to convict: they must be “satisfied and sure”,

i) that [Bentley] knew Craig had a gun and

ii) that he instigated or incited Craig to use it." (Trow p179)

The evidence adduced by the Prosecution to satisfy the jury on both points was linguistic.

Derek Bentley Case

Excerpt from "... and then ... Language Description and Author Attribution" by Malcolm Coulthard

For point i) it was observed that in his statement, which purported to give his unaided account of the night's events, Bentley had said "I did not know he was going to use the gun", (sentence 23).

In his summing up, the judge who, because of the importance of the case was the Lord Chief Justice, made great play with this sentence, telling the jury that its positioning in the narrative of events, before the time when there was a single policeman on the roof, combined with the choice of "the gun" (as opposed to "a gun") must imply that Bentley knew that Craig had a gun well before it was used.

In other words "the gun", given its position in the statement, must be taken to mean "the gun I already knew that Craig had".

Derek Bentley Case

Excerpt from "… and then … Language Description and Author Attribution" by Malcolm Coulthard

The evidence used to support point ii), that Bentley had instigated Craig to shoot, was from the police officers. In their written statements and in their verbal evidence in court, they asserted that Bentley had uttered the words "Let him have it, Chris" immediately before Craig had shot and killed the policeman.

As the judge emphasised, the strength of the linguistic evidence depended essentially on the credibility of the police officers who had remembered it recorded it, written it down later and then sworn to its accuracy.

When the case came to Appeal in 1998, one of the defence strategies was to challenge the reliability of Bentley's statement.

If they could throw doubt on the veracity of the police, they could mitigate the incriminating force of both the statement and the phrase "Let him have it", which Bentley, supported by Craig, had vehemently denied uttering.

Derek Bentley Case

Excerpt from "… and then … Language Description and Author Attribution" by Malcolm Coulthard

At the time of Bentley's arrest the police were allowed to collect verbal evidence from those accused of a crime in two ways:

- either by interview, when they were supposed to record contemporaneously, verbatim and in longhand, both their own questions and the replies they elicited,
- or by statement, when the accused was invited to write down, or, if s/he preferred, to dictate to a police officer, their version of events.

During statement-taking the police officers were supposed not to ask substantive questions.

At trial three police officers swore on oath that Bentley's statement was the product of unaided monologue dictation, whereas Bentley asserted that it was, in part at least, the product of dialogue, and that police questions and his replies to them had been reported as monologue.

There is no doubt that this procedure was sometimes used for producing statements.

Derek Bentley Case

Excerpt from "... and then ... Language Description and Author Attribution" by Malcolm Coulthard

There are many linguistic features which suggest that Bentley's statement is not, as claimed by the police, a verbatim record, see Coulthard (1993) for a detailed discussion; here we will focus only on evidence that the statement was indeed, at least in part, dialogue converted into monologue.

Firstly, the final four sentences of the statement

> (39) I knew we were going to break into the place. (40) I did not know what we were going to get - just anything that was going. (41) I did not have a gun and I did not know Chris had one until he shot. (42) I now know that the policeman in uniform that was shot is dead.

form some kind of meta-narrative whose presence and form are most easily explained as the result of a series of clarificatory questions about Bentley's knowledge at particular points in the narrative.

Derek Bentley Case

Excerpt from "… and then … Language Description and Author Attribution" by Malcolm Coulthard

… here are some negatives in Bentley's statement which have no … narrative justification.

… the most reasonable explanation for the negatives … is that, at this point in the statement-taking process, a policeman asked a clarificatory question to which the answer was negative and the whole sequence was then recoded and recorded as a negative statement by Bentley.

The fact that some of the statement may have been elicited in this way is of crucial importance in sentence (23):

(23) **I did not know** he was going to use the gun

Derek Bentley Case

Excerpt from "... and then ... Language Description and Author Attribution" by Malcolm Coulthard

This is the one singled out by the judge as incriminating.

This sentence would only make narrative sense if it were linked backwards or forwards to the use of a gun - in other words if it has been placed immediately preceding or following the report of a shot.

However, the actual context is:

> (22) We were there waiting for about ten minutes.
>
> (23) **I did not know** he was going to use the gun.
> (24) A plain clothes man climbed up the drainpipe and on to the roof.

If it is accepted that there were question/answer sequences underlying Bentley's statement, it follows that the logic and the sequencing of the information were not under his direct control.

Thus the placing of the reporting of some of the events must depend on a decision by the police questioner to ask his question at that point, rather than on Bentley's unaided reconstruction of the narrative sequence.

Derek Bentley Case

Excerpt from "… and then … Language Description and Author Attribution" by Malcolm Coulthard

Therefore, and crucially, this means that the inference drawn by the judge in his summing up about Bentley's prior knowledge of Craig's gun was totally unjustified -

if the sentence is the product of a response to a question, with its placing determined by the interrogating police officers,

there is no longer any conflict with Bentley's later denial "I did not know Chris had one [a gun] until he shot".

Nor is there any significance either to be attached to Bentley saying "the gun".

All interaction uses language loosely and co-operatively and so,

if the policeman had asked Bentley about "the gun",

Bentley would have assumed they both knew which gun they were talking about.

In that context the sensible interpretation would be 'the gun that had been used earlier that evening' and not 'the gun that was going to be used later' in the sequence of events that made up Bentley's own narrative of the evening.

Derek Bentley Case

Excerpt from "... and then ... Language Description and Author Attribution" by Malcolm Coulthard

Using corpus evidence

One of the marked features of Derek Bentley's confession is the frequent use of the word "then" in its temporal meaning - 11 occurrences in 588 words.

This may not, at first, seem at all remarkable given that Bentley is reporting a series of sequential events and that one of the obvious requirements of a witness statement is accuracy about time.

However, a cursory glance at a series of other witness statements showed that Bentley's usage of "then" was at the very least atypical, and thus a potential intrusion of a specific feature of policeman register deriving from a professional concern with the accurate recording of temporal sequence.

Derek Bentley Case

Excerpt from "… and then … Language Description and Author Attribution" by Malcolm Coulthard

Two small corpora were used to test this hypothesis,

the first composed of three ordinary witness statements, one from a woman involved in the Bentley case itself and two from men involved in another unrelated case,

totalling some 930 words of text,

the second composed of statements by three police officers, two of whom were involved in the Bentley case, the third in another unrelated case,

totalling some 2270 words.

Derek Bentley Case

Excerpt from "… and then … Language Description and Author Attribution" by Malcolm Coulthard

The comparative results were startling:

whereas in the ordinary witness statements there is only one occurrence,

"then" occurs 29 times in the police officers' statements,

that is an average of once every 78 words.

Thus, Bentley's usage of temporal "then",

once every 53 words,

groups his statement firmly with those produced by the police officers.

Derek Bentley Case

Excerpt from “… and then … Language Description and Author Attribution” by Malcolm Coulthard

In this case it was possible to check the findings from the ‘ordinary witness’ data against a reference corpus,

the Corpus of Spoken English, a subset of the COBUILD Bank of English, which, at that time, consisted of some 1.5 million words.

"Then" in all its meanings proved to occur a mere 3,164 times,

that is only once every 500 words,

which supported the representativeness of the witness data

and the claimed specialness of the data from the police and Bentley, (cf Fox 1993).

Derek Bentley Case

Excerpt from "... and then ... Language Description and Author Attribution" by Malcolm Coulthard

What was perhaps even more striking about the Bentley statement was the frequent post-positioning of the "then"s, as can be seen in the two sample sentences below, selected from a total of 7:

> Chris **then** jumped over and I followed.
>
> Chris **then** climbed up the drainpipe to the roof and I followed.

The opening phrases have an odd feel,

because not only do ordinary speakers use "then" much less frequently than policemen,

they also use it in a structurally different way.

For instance, in the COBUILD spoken data "then I" occurred ten times more frequently than "I then";

indeed the structure "I then" occurred a mere 9 times, in other words only once every 165,000 words.

Derek Bentley Case

Excerpt from "... and then ... Language Description and Author Attribution" by Malcolm Coulthard

By contrast the phrase occurs 3 times in Bentley's short statement,

once every 194 words,

a frequency almost a thousand times greater.

In addition, while the "I then" structure, as one might predict from the corpus data, did not occur at all in any of the three witness statements,

there were 9 occurrences in one single 980 word police statement,

as many as in the entire 1.5 million word spoken corpus.

Thus, the structure "I then" does appear to be a feature of policeman's (written) register.

Derek Bentley Case

Excerpt from "… and then … Language Description and Author Attribution" by Malcolm Coulthard

When we turn to look at yet another corpus,

the shorthand verbatim record of the oral evidence given in court during the trial of Bentley and Craig,

and choose one of the police officers at random,

we find him using the structure twice in successive sentences,

"shot him then between the eyes" and "he was then charged".

In Bentley's oral evidence there are also two occurrences of "then",

but this time the "then"s occur in the normal preposed position: "and then the other people moved off", "and then we came back up".

Even Mr. Cassels, one of the defence barristers, who might conceivably have been influenced by police reporting style, says "Then you".

Derek Bentley Case

Excerpt from "... and then ... Language Description and Author Attribution" by Malcolm Coulthard

Thus these examples, embedded in Bentley's statement, of the language of the police officers who had recorded it,

added support to Bentley's claim that it was a jointly authored document

and so both removed the incriminating significance of the phrase "I didn't know he was going to use the gun"

and undermined the credibility of the police officers on whose word depended the evidential value of the claimed-to-be remembered utterance "Let him have it Chris".

In August 1998, 46 years after the event, the then Lord Chief Justice, sitting with two senior colleagues, criticised his predecessor's summing-up and allowed the Appeal against conviction.

THANK YOU!