Where did we go wrong? A retrospective look at the British National Corpus

Lou Burnard, Humanities Computing Unit, Oxford University

Abstract

The British National Corpus (BNC) has been a major influence on the construction of language corpora during the last decade, if only as a major reference point. This corpus may be seen as the culmination of a research tradition going back to the one-million word Brown corpus of 1964, but its constitution and its industrial-scale production techniques look forward to a new world in which language-focussed engineering and software development are at the heart of the information society instead of lurking on its academic fringes.

This paper attempts to review the design and management issues and decisions taken during the construction of the BNC and to suggest what lessons have been learned over the last five years about how such corpus building exercises can most usefully be extended into the new century.

1. What, exactly, is the BNC?

The British National Corpus (BNC) is a 100 million word corpus of modern British English, originally produced by a consortium of dictionary publishers and academic researchers in 1990-1994. The Consortium brought together as members dictionary publishers OUP, Longman, and Chambers, research centres at the Universities of Lancaster and Oxford along with the British Library's now abolished Centre for Research and Development. The project was originally funded under the Joint Framework for Information Technology, a British Government initiative designed to facilitate academic-industrial co-operation in the production of what were regarded as 'pre-competitive' resources, whereby the Department of Trade and Industry provided 50 percent funding to commercial partners, and the Science and Engineering Research Council funded 100 percent of the academics' costs.

The nineties have been called many things in social history: as far as computing facilities are concerned however, I suggest that an appropriate epithet might well be *neotenous*. It is salutary to remember that in computer magazines of the early nineties, the big debate was about the relative merits of word processors WordPerfect release 5 and WinWord (an ancestor of the now ubiquitous Microsoft Word). On your desktop, if you were a reasonably well funded academic, you might have a 'personal computer' with a fast Intel 386 processor, and as much as 50 Mb of diskspace — just about enough to run Microsoft's new-fangled Windows 3.1 operating system. But your real computing work would be done in your laboratory or at your centralised computing service, where you would probably have shared use of a Unix system of some kind or a VAX minicomputer. This was also a period in which a few people were starting to talk about a new hypertext concept called the World Wide Web, a few of whom might even have tried an impressive new interface programme called Mosaic...

The art of corpus building was however already well understood in the nineties, at least by its European practitioners. "Corpus are becoming mainstream" declared Leech, with palpable surprise, in the preface to the ICAME proceedings volume of 1990. We may discern three intellectual currents or differences of emphasis already becoming clear at this period: the traditional school initiated by the Brown Corpus, institutionalised in LOB, and perpetuated through ICAME; the Birmingham school, which had been building up ever larger collections of textual material as part of the COBUILD project throughout the late eighties[1]; and the

[1]For a summary of this highly influential work, prefiguring that of the BNC in many regards, see Renouf 1986, and the many publications of its intellectual centre, J. McH. Sinclair, e.g. Sinclair 1987

American view most famously expressed by Mitch Marcus as "there's no data like more data". The locale in which these traditions most visibly began to combine into a new form was in computer aided lexicography, partly as a consequence of the availability of computer-held representations of traditionally organised dictionaries, such as Longman's Dictionary of Contemporary English, and of course the computerization of the Oxford English Dictionary itself, partly as a result of an upsurge of interest amongst the computational linguistics community (see for example Atkins92).

At the same time, the early 90s were an exciting period for synergy in research applications of IT. 'Humanities Computing' and Computational Linguistics were pulling together in their first (and to date only) joint success, the establishment of Text Encoding standards appropriate to the dawning digital age.[2] The term *language engineering* was being used to describe not a dubious kind of social policy, but a sexy new sort of technology. It is in this context that we should place the fact that production of the BNC was funded over three years, with a budget of over GBP 1.5 million.

The project came into being through an unusual coincidence of interests amongst lexicographic publishers, government, and researchers. Amongst the publishers, Oxford University Press and Longman were at that time beginning to wake up to the possible benefits of corpus use in this field. One should point also to the success of the Collins COBUILD dictionaries, (first published in 1987, and probably the first major language-learner dictionary whole-heartedly to embrace corpus principles) as a vital motivating factor for rival publishers OUP and Longman. For the government, a key factor was a desire to stimulate a UK language engineering industry in the climate of expanded interest in this field in Europe. For researchers at Oxford and Lancaster, this unlikely synergy was a golden opportunity to push further the boundaries of corpus construction, as further discussed below. And for the British Library, the corpus was one of a number of exploratory projects being set up to experiment with new media at the beginning of the age of the digital library (for other examples, see the essays in Carpenter 1998)

The stated goals of the BNC project were quite explicit: it would create a language corpus at least an order of magnitude bigger than any freely available hitherto[3]The new corpus would be synchronic and contemporary and it would comprise a range of samples from the full range of British English language production, both spoken and written. Considerable debate and discussion focussed on the notion of sampling, and in particular of corpus design. Unlike some other collections of language data then popular, the BNC would be of avowedly non-opportunistic design. In order to make the corpus generally applicable, it would contain automatically generated word class annotation, and it would also include very detailed contextual information. These three features, together with its general availability and large size, would make the BNC unique amongst available collections of language data, and would also justify the 'national' part of its title (originally included simply in recognition of the fact that the project was partly government funded).

Unstated, but clearly implicit in the project design, were other goals. For the commercial partners, the major reason for their substantial investment of time and money was of course the production of better ELT dictionaries, plus, perhaps some regaining of competitive position by the authoritative nature of the resulting corpus. For the academic partners, an unstated goal was to provide a new model for the development of corpora within the emerging European language industries, and to put to the test emerging ideas about standardization of

[2]The introduction to Zampolli 1994 makes this connexion explicit.

[3]Though incontestably larger, the Bank of English corpus developed as part of the Cobuild project was not originally designed for distribution or use by anyone outside that project; to this day, IPR and other restrictions have effectively limited access to it by the research community at large.

encoding and text representation and documentation. But over-riding all there was the simple desire to build a really big corpus!

2. Organization of the Project

An interesting and unanticipated consequence of the academic - industrial co-operation was the need for some accomodation between the academic desire for perfection and the commercial imperatives of delivering a pre-defined product on time and not too far over budget (see further Burnard 1999). In setting up an industrial scale text production system, the project found itself inevitably making compromises in both design and execution. The production line itself, dubbed by project manager Jeremy Clear the *BNC Sausage machine* is shown in the following figure:

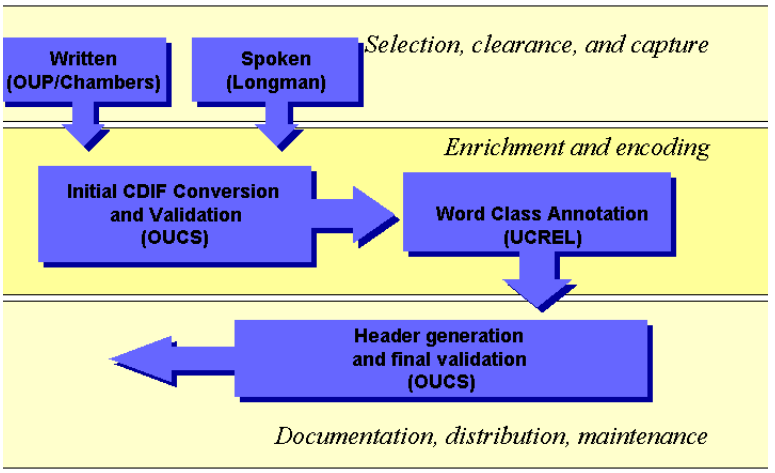

Figure 1. The BNC Sausage Machine

As this figure demonstrates, production of different types of material was shared out amongst a number of different agencies: Longman focussed on the collection and transcription of spoken materials, and OUP on the transcription of written materials, using a mixture of OCR, rekeying, and processing of materials already in digital form. Conversion of all materials to a single consistent format and validation of its structure was carried out at OUCS, which also maintained a database of contextual and workflow information. Linguistic annotation of the material was carried out at Lancaster, using the well-established CLAWS tagger (discussed below and in Garside 1996, and the resulting texts then combined with standard metadata descriptions extracted from the database to form a single document conformant (as far as these were so far published) to the recommendations of the Text Encoding Iniative (Sperberg-McQueen 1994).

As might be expected, the rate with which the sausage machine turned was far from constant over the life of the project, and there were inevitably temporary blockages and hold ups. The figure below demonstrates through-put (in millions of words per quarter) over the life time of the project: Through-put is shown separately for each of: material received from the data preparation agency; texts validated against the DTD; and texts annotated with POS codes.

The work of developing the corpus was shared out amongst five task groups, on which staff from each of the consortium members participated to varying extents. These task groups and their responsibilities may be summarised as follows:

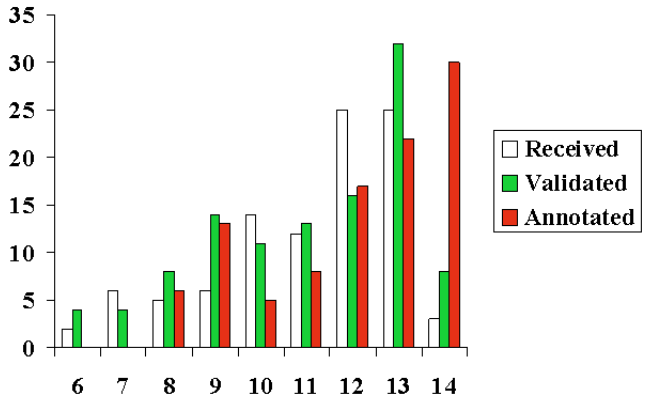

Figure 2. BNC Through-put

permissions design and implementation of a standard permissions letter for use with all those owning IPR in the materials to be included in the corpus;

selection, design criteria definition of the range of text types to be included in the corpus and of their target proportions;

enrichment and annotation implementation of linguistic and contextual annotation of the corpus texts;

encoding and markup definition of the markup scheme to be applied in the final reference form of the corpus, and of procedures for mapping to it from a variety of data capture formats;

retrieval software definition and implementation of simple retrieval software able to make use of the detailed corpus encoding.

Each of these topics is further discussed in the following sections.

2.1. Permissions Issues

As noted above, the BNC was the first corpus of its size to be made widely available. This was possible largely because of the work done by this task group in successfully defining standard forms of agreement, between rights owners and the Consortium on the one hand, and between corpus users and the Consortium on the other. IPR owners were requested to give permission for the inclusion of their materials in the corpus free of charge, and shown the standard licence agreement which is still used today. Acceptance of this arrangement was perhaps to some extent facilitated by the relative novelty of the concept and the prestige attached to the project; however by no means every rights owner approached was immediately ready to assign rights to use digital versions of their material for linguistic research purpose indefinitely and free of charge. Some chose to avoid committing themselves at all, and others refused any non-paying arrangements.

Two specific problems attached to permissions issues relating to the spoken materials. Because participants had been assured that their identies would be kept secret, much effort was put into pondering how best to anonymise the their contributions, without unduly compromising their linguistic usefulness. Specific references to named persons were in many cases removed; the option of replacing them by alternative (but linguistically similar) names was briefly considered but felt to be impractical.

A more embarassing problem derives from the fact that participants in the demographically sampled part of the corpus had been asked (and had therefore given) permission only for inclusion of transcribed versions of their speech, not for inclusion of the speech itself. While such permission could in principle be sought again from the original respondents, the effectiveness of the anonymization procedures used now makes this a rather difficult task.

Two additional factors affected the willingness of IPR owners to donate materials: firstly, that no complete texts were to be included; secondly, that there was no intention of commercially exploiting or distributing the corpus materials themselves. This did not however preclude commercial usage of derived products, created as a consequence of access to the corpus. This distinction, made explicit in the standard User Licence, is obviously essential both to the continued availability of the corpus for research purposes, and to its continued usefulness in the commercial sector, for example as a testbed for language products from humble spelling correction software to sophisticated translation memories. To emphasize the non-commercial basis on which the corpus itself was to be distributed, one of the academic members of the consortium, OUCS, was appointed sole agent for licensing its use, reporting any dubious cases to the Consortium itself. Initially restricted to the EU, distribution of the corpus outside Europe was finally permitted in 1998.

2.2. Design Criteria

I referred above to the BNC's "non-opportunistic design". A sense of the historical context is also perhaps helpful to understand the singling out of this aspect of the design as noteworthy. During the mid-nineties, although textual materials of all kinds were increasingly being prepared in digital form as a precursor to their appearance in print, the notion that the digital form might itself be of value was not at all widespread. Moreover, digitization in those pre-e-commerce days was far from uniform either in coverage or in format. As a consequence, there was a natural tendency in the research community to snap up such unconsidered trifles of electronic text as were available without considering too deeply their status with respect to the language as a whole. Because, to take one notorious example, large quantities of the *Wall Street Journal* were widely available in digital form, there was a danger that the specific register typified by that newspaper would increasingly serve as a basis for computationally-derived linguistic generalisations about the whole language.

As a corrective, therefore, the BNC project established at its outset the goal of sampling materials from across the language with respect to explicit design criteria rather than simply their contingent availability in machine-readable form. These criteria (usefully summarized in Atkins 1992) defined a specific range of text characteristics and target proportions for the material to be collected. The goal of the BNC was to make it possible to say something about language in general. But is language that which is *received* (read and heard) or that which is *produced* (written and spoken)? As good Anglo-Saxon pragmatists, the designers of the BNC chose to ignore this classic Saussurian dichotomy by attempting to take account of both perspectives.

The objective was to define a stratified sample according to stated criteria, so that while no-one could reasonably claim that the corpus was statistically representative of the whole language in terms either of production or reception, at least the corpus would represent the degree of variability known to exist along certain specific dimensions, such as mode of production (speech or writing); medium (book, newspaper, etc.); domain (imaginative, scientific, leisure etc.); social context (formal, informal, business, etc.) and so on.

This is not the place to rehearse in detail the motivations for the text classification scheme adopted by the BNC[4]. For example, spoken texts may be characterized by age, sex, or social class (of respondent, not speaker), by the domain, region, or type of speech captured; written texts may also be characterized by author age, sex, type, by audience, circulation, status, and (as noted above) by medium or domain. Some of these categories were regarded as *selection criteria*, i.e. the domain of values for this category was predefined, and a target proportion identified for each; while others were regarded as *descriptive criteria*, i.e. while no particular target was set for the proportion of material of a particular type, other things being equal, attempts would be made to maximize variability within such categories. It should be stressed that the purpose of noting these variables was to improve coverage, *not* to facilitate accessibility or subsetting of the corpus.

Inevitably, the design goals of the project had to be tempered by the realities of economic life. A rough guess suggests that the cost of collecting and transcribing in electronic form one million words of naturally occurring speech is at least 10 times higher than the cost of adding another million words of newspaper text: the proportion of written to spoken material in the BNC is thus 10:1, even though most people would suggest that if speech and writing are of equal significance in the language, they should therefore be present in equal amounts in the corpus. Within the spoken corpus, an attempt is made to represent equally the production of different speech types (in the context-governed part) and its reception (in the demographically sampled part).

Similarly pragmatic concerns lead to the predominance within the written part of the corpus of published books and periodical. However, while text that is published in the form of books, magazines, etc., may not be representative of the totality of written language that is produced, (since writing for publication is a comparatively specialized activity in which few people engage), it is obviously representative of the written language that most people receive. In addition, it should be noted that significant amounts of other material (notably unpublished materials such as letters or gray literature) are also included. And even within a readily accessible text-type such as newspapers, care was taken to sample both broadsheet and tabloid varieties, both national and regional in such a way that the readily available (national broadsheet) variety did not drown out the other, less readily found, variants.

In its final form, the BNC World Edition contains 4054 texts and occupies (including SGML markup) 1,508,392 Kbytes, or about 1.5 Gb. In total, it comprises just over 100 million orthographic words (specifically, 100,467,090), but the number of w-units (POS-tagged items) is slightly less: 97,619,934. The total number of s-units identified by CLAWS is just over 6 million (6,053,093). The following table shows the breakdown 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)

Table 1. Composition of the BNC World Edition

Text type	Texts	Kbytes	W-units	S-units	percent
Spoken demographic	153	4206058	4.30	610563	10.08

[4]These are exhaustively discussed in e.g. Atkins 1992 for the written material, and Crowdy 1995 for the spoken material; discussion and detailed tables for each classification are also provided in the *BNC User Reference Guide* (Burnard 1995, revised 2000).

Spoken context-governed	757	6135671	6.28	428558	7.07
All Spoken	910	10341729	10.58	1039121	17.78
Written books and periodicals	2688	78580018	80.49	4403803	72.75
Written-to-be-spoken	35	1324480	1.35	120153	1.98
Written miscellaneous	421	7373707	7.55	490016	8.09
All Written	3144	87278205	89.39	5013972	82.82

Within the written part of the corpus, target proportions were defined for each of a range of types of media, and subject matter. Here for example are the counts for written domain:

Table 2. Written domain

Domain	texts	w-units	%	s-units	%
Applied science	370	7104635	8.14	357067	7.12
Arts	261	6520634	7.47	321442	6.41
Belief and thought	146	3007244	3.44	151418	3.01
Commerce and finance	295	7257542	8.31	382717	7.63
Imaginative	477	16377726	18.76	1356458	27.05
Leisure	438	12187946	13.96	760722	15.17
Natural and pure science	146	3784273	4.33	183466	3.65
Social science	527	13906182	15.93	700122	13.96
World affairs	484	17132023	19.62	800560	15.96

The spoken part of the corpus is itself divided into two. Approximately half of it is composed of informal conversation recorded by approximately 200 volunteers recruited for the project by a market research agency and forming a balanced sample with respect to age, gender, geographical area, and social class. This sampling method reflects the demographic distribution of spoken language, but (because of its small size) would have excluded from the corpus much linguistically-significant variation due to context. To compensate for this, the other half of the spoken corpus consists of speech recorded in each of a large range of predefined situations (for example public and semi-public meetings, professional interviews, formal and semi-formal proceedings in academia, business, or leisure contexts).

In retrospect, some of these classifications were poorly defined and many of them were only partially or unreliably populated. Pressures of production and lack of ready information in some cases seriously affected the accuracy and consistency with which these variables were actually recorded in the text headers: in some cases (author ethnic origin for example) the concepts recorded were not very well defined. Even such a seemingly neutral concept as dating is not unproblematic for written text — are we talking about date of the copy used or of the first publication? Similarly, when we talk of “Author age” do we mean age at the time the book was published, or when it was printed?

Of course, corpora before the BNC had been designed according to similar methods, though perhaps not on such a scale. In general, however, the metadata associated with such corpora had been regarded as something distinct from the corpus itself, to be sought out by the curious in the ‘manual of information to accompany’ the corpus. One innovation due to the Text Encoding Initiative, and adopted by the BNC, was the idea of an integrated *header*term>, attached to each text file in the corpus, and using the same formalism. This header contains information identifying and classifying each text, as well as additional

specifics such as demographic data about the speakers, and housekeeping information about the size, update status, etc. Again following the TEI, the BNC factors out all common date (such as documentation and definition of the classification codes used) into a header file applicable to the whole corpus, retaining within each text header only the specific codes applicable to that text.[5]

During production, however, classificatory and other metadata was naturally gathered as part of the text capture process by the different data capture agencies mentioned above and stored locally before it was integrated within the OUCS database from which the TEI headers were generated. With the best will in the world, it was therefore difficult to avoid inconsistencies in the way metadata was captured, and hence to ensure that it was uniformly reliable when combined. This is a problem which we did not have leisure to address during production of BNC1.

2.3. Annotation

Word tagging in the BNC was performed automatically, using CLAWS4, an automatic tagger developed at Lancaster University from the CLAWS1 tagger originally produced to perform a similar task on the one million LOB Corpus. The system is described more fully in Leech 1994; its theory and practice are explored in Garside 1997, and full technical documentation of its usage with the BNC is provided in the Manual which accompanies the BNC World Edition (Leech 2000).

CLAWS4 is a *hybrid* tagger, employing a mixture of probabilistic and non-probabilistic techniques. It assigns a part-of-speech code (or sometimes two codes) to a word as a result of four main processes:

- tokenization into words (usually marked by spaces) and orthographic sentences (usually marked by punctuation); enclitic verbs (such as *'ll* or *'s*), and negative contractions (such as *n't*) are regarded as special cases, as are some common merged forms such as *dunno* (which is tokenized as "do + n't + know"
- initial POS code assignment: all the POS codes which might be assigned to a token are retrieved, either by lookup from a 50,000 word lexicon, or by application of some simple morphological procedures; where more than one code is assigned to the word, the relative probability for each code is also provided by the lexicon look-up or other procedures. Probabilities are also adjusted on the basis of word-position within the sentence.
- disambiguation or code selection is then applied, using a technique known as *Viterbi alignment* which chooses the probabilities associated with each code to determine the most likely path through a sequence of ambiguous codes, in rather the same way as the text messaging applications found on many current mobile phones. At the end of this stage, the possible codes are ranked in descending probability for each word in its context
- idiom tagging is a further refinement of the procedure, in which groups of words and their tags are matched against predefined idiomatic templates, resembling finite-state networks.

With these procedures, CLAWS was able to achieve over 95% accuracy (i.e. lack of indeterminacy) in assigning POS codes to any word in the corpus. To improve on this, the Lancaster team developed further the basic ideas of 'idiom tagging', using a template tagger which could be taught more sophisticated contextual rules, in part derived by semi-automatic procedures from a sample set of texts which had previously been manually disambiguated. This process is further described in the Reference Manual cited.

[5] For further description of the way TEI Headers are used by the BNC see Dunlop 1995

2.4. Encoding

The markup scheme used by the BNC was defined at the same time as the Text Encoding Initiative's work was being done (and to some extent by the same people); the two schemes are thus unsurprisingly close, though there are differences. Since this SGML-based scheme has been so widely taken up and is well documented elsewhere, we do not discuss it further here.

As some indication of the extent and nature of the markup in the BNC, here is the start of a typical written text:

```
<text complete=Y decls='CN000 HN001 QN000 SN000'>
<div1 complete=Y org=SEQ>
<head type=MAIN>
<s n=001>
<w NP0>CAMRA <w NN1>FACT <w NN1>SHEET
<w AT0>No <w CRD>1 </head>
<head r=it type=SUB>
<s n=002><w AVQ>How <w NN1>beer <w VBZ>is
<w AJ0-VVN>brewed </head>
<p><s n=003>
<w NN1>Beer <w VVZ>seems <w DT0>such
<w AT0>a <w AJ0>simple <w NN1>drink <w CJT>that
<w PNP>we <w VVB>tend <w TO0>to <w VVI>take
<w PNP>it <w CJS-PRP>for <w VVD-VVN>granted
<c PUN>.
```

The start of each 'word' identified by CLAWS is marked by an SGML <w> tag, which also contains the POS code or codes allocated to it. The start of each 'sentence' is similarly marked by an SGML <s> element, carrying the sentence number of this sentence within the text. SGML elements such as <head> and <p> are used to mark larger structural components of the text such as headings and paragraphs.

The User Reference Guide (Burnard 1995) delivered with the corpus contains a detailed discussion of the scope and significance of this markup system. As the above example shows, it relies on the minimization features of SGML to reduce the quantity of markup: in XML, which lacks this facility, the last few words of the above example would read:

```
... <w type="VVI">take</w> <w type="PNP">it</w>
<w type="CJS-PRP">for</w> <w type="VVD-VVN">granted</w><c type="PUN">.
</s>
```

representing an additional overhead of approximately 10 bytes per token, or approximately 1000 Mb for the whole corpus — not an attractive thought, even with plummeting disk space prices and increasingly effective compression algorithms.

In marking up the spoken part of the corpus, many different technical issues had to be addressed. As noted above, this was the first time SGML markup of transcribed speech on such a scale had been attempted. The transcription itself was carried out by staff who were not linguistically trained (but who were however familiar with the regional variation being transcribed — staff recruited in Essex for example were not required to transcribe material recorded in Northern Ireland). Transcribers added a minimal (non-SGML) kind of markup to the text, which was then normalized, converted to SGML, and validated by special purpose software (see further Burnage 1993). The markup scheme made explicit a number of features, including changes of speaker and quite detailed overlap; the words used, as perceived by the transcriber; indications of false starts, truncation, uncertainty; some performance features e.g. pausing, stage directions etc. In addition, of course, detailed demographic and other

information about each speaker and each speech context was recorded in the appropriate part of the Header, where this was available.

Here is a sample from a transcribed spoken text:

```
<u who=PS04Y>
<s n=01296><w ITJ>Mm <pause> <w ITJ>yes <pause dur=7>
<w PNP>I <w VVD>told <w NP0>Paul <pause>
<w CJT>that <w PNP>he <w VM0>can <w VVI>bring
<w AT0>a <w NN1>lady <w AVP>up <pause> <w PRP>at
<w NN1>Christmas-time<c PUN>.</u>
<u who=PS04U>
<s n=01297><w VBZ>Is <w PNP>he <w XX0>not
<w VVG>going <w AV0>home <w AV0>then<c PUN>?</u>
<u who=PS04Y>
<s n=01298><w ITJ>No <pause dur=8> <w CJC>and
<w UNC>erm <pause dur=7> <w PNP>I<w VBB>'m
<w VVG>leaving <w AT0>a <w NN1>turkey <w PRP>in
<w AT0>the <w NN1>freezer
<event desc="kettle boils">
<s n=01299><w NP0>Paul <w VBZ>is <w AV0>quite
<w AJ0>good <w PRP>at <w NN1-VVG>cooking <pause>
<w AJ0>standard <w NN1>cooking<c PUN>.</u>
```

Words and sentences are tagged as in the written example above. However, sentences are now grouped into *utterances*, marked by the SGML `<u>` element, each representing an unbroken stretch of speech, and containing withing its start tag a code (such as PS04Y) which acts as a key to access the more detailed information about the speaker recorded in the TEI Header for this text. Note also the `<pause>` and `<event>` elements used to mark paralinguistic features of the transcribed speech.

As can readily be seen in the above example, the intention of the transcribers was to provide a version of the speech which was closer to writing than to unmediated audio signal. Thus, the spelling of filled pauses such as *erm* or *mmm* is normalised, and there is even use of conventional punctuation to mark intonation patterns interpreted as questions. For more discussion of the rationale behind this and other aspects of the speech transcription see Crowdy 1994.

2.5. Software and distribution

In 1994, it was not entirely obvious how one should distribute a corpus the size of the BNC on a not-for-profit basis. Low-cost options such as anonymous ftp seemed precluded by the scale of the data. Our initial policy was to distribute the text compressed to the extent that it would fit on a set of three CDs, together with some simple software system which could be installed by suitably skilled personnel to provide departmental access over a network to the local copy of the corpus. Development of such a software system was undertaken, with the aid of additional funding from the British Library, during the last year of the project. The software, named SARA (for SGML Aware Retrieval Application), has since been further developed as discussed below, but remains essentially unchanged in its mode of usage: in particular, it remains a client-server application, not well suited to small memory-intensive operations.

It was always intended that access to the BNC should not be contingent on the use of any particular software — this was after all the main rationale behind the use of the international standard SGML as a means of encoding the corpus, rather than a system tailored to any particular software tool. To demonstrate this software independence, a small (two million word) extract from the corpus was also produced which included with it on a single

CD customised versions of three other software systems: Mike Scott's WordSmith, Oliver Mason's Qwick, and the Corpus Workbench software developed at Stuttgart, as well as an enhanced version of SARA. This 'BNC Sampler' proved very popular in introducing the BNC to a wider audience; it also clearly demonstrated that there was a large demand for such corpora in standalone computing environments, such as individual desktop machines.

As noted above, the BNC dates from the pre-World Wide Web era. [6] However, within a year of its publication, it was apparent that web access would be the ideal way of making it available, if only because this would enable us to provide a service to researchers outside the European Union, who were still at this time unable to obtain copies of the corpus itself because of licencing restrictions. The British Library generously offered the project a server for this purpose, and a simple web interface to the corpus was developed. This service, still available at the address http://sara.natcorp.ox.ac.uk allows anyone to perform basic searches of the corpus, with a restricted range of display options; those wishing for more sophisticated facilities can also download a copy of the SARA client program to access the same server: a small registration fee is charged for continued use of the service beyond an initial trial period.

To complement this service, and in response to the demand for help in using the BNC from the language teaching community, a detailed tutorial guide (Aston 1999) was written, introducing the various facilities of the software in the form of focussed and linguistically-motivated exercises. The Online service remains very popular, receiving several thousand queries each month.

Further evidence of the benefits of the software-independent approach taken in designing the corpus is provided by the success of the BNCweb project at the University of Zurich (Lehmann 1999), which has developed an entirely web-based approach to searching the BNC using the SARA server as a back end together with a database of associated word-frequency data.

3. The BNC World Edition

The much-delayed BNC World Edition (also known as BNC2) was published in December 1999, five years after the first appearance of the BNC. A small number (less than 50) texts for which world rights could not be obtained were removed from the corpus so that it could, at last, be distributed worldwide.

Desirable though it might be, the scale of the BNC precludes any complete proof reading of it. The BNC's function as a snapshot of British English in the mid-nineties also precludes adding more material to it. Nevertheless, we were able to make several revisions and corrections, described briefly below. In preparing this new edition, we were also able to catch up with the standards established after (and to some extent by) the BNC itself, and to provide a new enhanced version of SARA.

Despite the growing popularity of XML the BNC World Edition is still in SGML, for reasons referred to above, but the DTD it uses is now TEI-conformant, and there is a section in the manual which defines formally its relationship with both the published TEI specification, and its derivative Corpus Encoding Scheme (Ide 1996). In practice, the differences are very slight — a few elements have different names, and the content model used by the BNC is simpler and more restrictive than that of CES. Although the text remain in SGML, the headers of the World Edition are now expressed in XML, which means that they can be processed by standard XML software to create for example an XML database.

[6] The phrase *world wide web* in fact appears only twice in the corpus, in both cases as part of a brief exchange about the feasibility of publicizing the Leeds United football club which occurred on an email discussion list in January 1994. The most frequent collocates for the word *web* in the corpus are *spider*, *tangled*, *complex*, and *seamless*. In this respect at least the BNC is definitely no longer an accurate reflection of the English language.

Trying to correct the errors in the BNC is not unlike the task of sweeping a beach clear of sand, as imagined by the Walrus and the Carpenter:

> "If seven maids with seven mops swept it for half a year
> Do you suppose," the Walrus said, "that they would get it clear?"
> "I doubt it," said the Carpenter and shed a heavy tear.

There is a sense in which any transcription of spoken text is inevitably indeterminate. Even for written texts deciding what counts as an error is not always obvious: mis-spelled words do appear in published material, and should therefore also be expected to appear in a corpus. Where corrections have been made during the process of corpus construction, they are sometimes noted in the markup in such a way as to preserve both the original error and its correction: this provides some indication at least of the kinds of error likely to be encountered. However, it is impossible reliably to assess the extent of such errors, nor precisely to locate their origin, because of the varied processes carried out on the source texts. In principle, it is impossible to distinguish an error introduced by (for example) inaccurate OCR software from an error which was present in the original, without doing an exact proof reading of the text against its original source[7]; the use of automatic spelling-error detection software also somewhat muddies the water.

One kind of systematic correction is however possible, and has been applied. In part because of the availability of the BNC Sampler, it was possible to improve greatly the rules used by CLAWS, and thus to significantly reduce both the error rate and the degree of indeterminacy in the POS codes for BNC world. This work, carried out at Lancaster with funding from the Engineering and Physical Sciences Research Council (Research Grant No. GR/F 99847), is described in detail in the Manual supplied with the corpus (Leech 2000), which estimates that the error rate in the whole corpus following the automatic procedures applied is now reduced to approximately 1.15 percent of all words, while the proportion of ambiguous codes is now reduced to approximately 3.75 per cent[8]

At the same time, a number of semi-systematic errors were fixed. These ranged from duplicate or wrongly labelled texts, to a complete check of the demographic data associated with the speakers in each text, which had been found to contain many errors in BNC1. In addition, we were able include the results of a systematic investigation of the text classifications carried out at Lancaster (reported in Lee 2001); this means that each text now carries not only a somewhat more reliable version of its original classification criteria, but also a completely new classification carried out in terms of a more delicate taxonomy defined by Lee for the corpus. Similarly, whereas in BNC-1 a rather unsystematic method had been employed to associate descriptive topic keywords with each text, in the new version, each written text has additionally been given the set of descriptive keywords associated with it in standard library catalogues.[9] The typos, however remain... and will continue to do so!

The new version of SARA distributed with the BNC World has five important new features, four of them suggested by user feedback on the original version. It is now possible to define subcorpora, using the improved classification codes for example, though this is not the only method. It is possible to carry out proper collocation analyses: finding (for example) all the words that collocate with a given term. It is possible to perform lemmatized searches, for example finding all morphologically related forms of a given word. And perhaps most important, the new version of SARA can be used with any TEI conformant corpus. More

[7] Such a task would however be feasible, since the original paper sources for the majority of the written parts of the corpus is still preserved at OUCS

[8] These estimates are derived from manual inspection of a 50,000 word sample taken from the whole corpus, as further discussed in the Tagging Manual cited.

[9] These were obtained from the UK joint COPAC for the bulk of the written published material

information on each of these facilities, together with practical information on how to install and use the new version is available from the website http://www.hcu.ox.ac.uk/SARA.

The fifth new feature of this release of the BNC is a consequence of the changing technical environment in which the corpus is now used. It is simply that the corpus can now be installed at low cost for personal use on a single standalone workstation running any version of the Windows operating system. This continues the trend initiated by the development of the BNC Online service towards making the corpus more accessible to a wider community of users.

4. What lessons have we learned?

Everyone knows you should research the market before distributing any kind of project, especially one with the level of initial investment needed by the BNC. But, as with some other things that everyone knows, this common-sense wisdom turns out to have been somewhat misleading in the case of the BNC. When the original project partners discussed the likely market for copies of the BNC, it seemed quite clear who and how small it would be. In the mid-nineties, it was obvious that only a specialist research community, with a clear focus on Natural Language Processing, and of course the research and development departments of businesses engaged in NLP or in lexicography would be in the least interested in a 100 million word collection of English in what was then still called machine-readable form. Both the rights framework for distribution of copies of the corpus and the methods of distribution chosen clearly reflect this 'obvious' model: the licence which all would-be purchasers must sign (in duplicate) for example talks about the licensees's "research group" and is quite belligerent about the need to monitor networked usage of the corpus within an institution — but nowhere entertains the notion that an individual might buy a copy for their own personal use, or for use with a group of students.

In fact however, we rapidly discovered that the market was both much larger, and quite different in nature. The major users of the BNC turn out to be people working in applied linguistics, not computational linguistics, and in particular those concerned with language learning and teaching. Their computational expertise is rather less than expected, their enthusiasms more wide-ranging. They include not only computational linguists and NLP researchers but also the full range of applied linguists, cultural historians and even language learners.

In retrospect, the BNC project also had the same technological blind spots as others at the time. Curiously, we did not expect the success of the XML revolution! So we wasted time in format conversion and compromises. Equally, because we did not foresee standalone computers running at 1 Ghz with 20 gigabyte disks as standard home equipment, we did not anticipate that it might one day be feasible to store the digital audio version of the texts we transcribed along with their transcriptions. Consequently, we never even considered whether it would be useful to try to get rights to distribute the digital audio, and our software development efforts focussed on developing a client/server application, a system predicated on the assumption that BNC usage would be characterized by a single shared computing resource, with many entry points, rather than by the massive duplication of standalone machines.

What other opportunities did we miss? In the original design, there is a clearly discernible shift from the notion of 'Representativeness' to the idea of the BNC as a *fonds*: a source of specialist corpora. From being a sample of the whole of language, the BNC was rapidly re-positioned as a repository of language variety. This was in retrospect a sensible repositioning; a more divers collection of materials than the BNC is hard to imagine. Handling this diversity effectively however requires a clearer and better agreed taxonomy of text types than currently

exists, and better access facilities for subcorpora. The BNC World edition tries to address this need by expanding the provision of classificatory data attached to each text, and the encoding scheme adopted certainly allows for the addition of any number of arbitrary classifications for each text (or, indeed, for each textual component) but there remains the disagreeable necessity of first defining and then applying such classications by hand in a consistent and defensible manner. A rapid scan of most corpus related discussion lists shows that close to the top of most frequently asked question lists is a question of the form "I am looking for a collection of texts of type X" (recent values for X I have noticed include doctor-patient interaction, legal debate, arguments, flirtation...); in almost every case, the answer to such a request is "There is some, somewhere in the BNC, but it's up to you to find it...".

Clearly, the design of the BNC entirely missed the opportunity to set up a grand monitor corpus, one which could watch the river of language flow and change across time. It is a rather depressing thought that linguists of this century may continue to study the language of the nineties for as long as those of the preceding one were constrained to study that of the sixties. It would be interesting, of course, to build a series of BNC-like corpora at regular intervals, say every decade, if only there were an unlimited supply of funding for such an enterprise. Instead, however, we will have a different kind of large scale corpus of language production at our disposal for at least the foreseeable future. How best to manage the diversity and unpredictability of the Web as our future source of linguistic information is another, and quite different, story.

5. Works cited

Aston, G. and Burnard, L. (1998), *The BNC handbook: exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

Atkins, B.T.S., Clear, J., and Ostler, N. (1992), 'Corpus Design Criteria', *Literary and Linguistic Computing*, 7, pp. 1-16.

Atkins, B.T.S., and Zampolli, A. (1994) eds. *Computational Approaches to the Lexicon*, Oxford: Oxford University Press.

Burnage, G. and Dunlop, D. (1992) 'Encoding the British National Corpus' in Aarts et al, eds. *English language corpora: design, analysis and exploitation* Amsterdam: Rodopi, pp 79-95

Burnard, L. (1995) ed. *Users' Reference Guide for the British National Corpus, version 1.0*. Oxford: Oxford University Computing Services.

Burnard, L. (1999) 'Using SGML for linguistic analysis: the case of the BNC' in *Markup languages theory and practice*. I.2 pp. 31-51. Cambridge, Mass: MIT Press. Also published in *Maschinelle Verarbeitung altdeitscher Texte V*, pp 53-72. Tuebingen: Max Niemeyer, 2001.

Carpenter, L., Shaw, S., and Prescott, A. (1998), eds. *Towards the Digital Library: the British Library's Initiatives for Access programme* London: British Library.

Clear, J. H. (1993) 'The British National Corpus' in Delany, P. and Landow, G., ed. *The Digital Word : text-based computing in the humanities*. Cambridge (Mass), MIT Press, pp. 163-187.

Crowdy, S. (1994) 'Spoken Corpus Transcription' *Literary & Linguistic Computing* 9:1 (1994), pp. 25-28.

Crowdy, S. (1995) 'The BNC spoken corpus' in Leech, G., Myers, G. and Thomas, J., eds. *Spoken English on computer: transcription, mark-up and application* Harlow: Longman, pp. 224-235.

Dunlop, D. (1995) 'Practical considerations in the use of TEI headers in large corpora' in Ide, N. and Veronis, J. eds. *Text Encoding Initiative: background and context*. Dordrecht: Kluwer, pp 85-98.

Garside, R. (1995) 'Grammatical tagging of the spoken part of the British National Corpus: a progress report' in Leech, G., Myers, G. and Thomas, J. , eds. *Spoken English on computer: transcription, mark-up and application* Harlow: Longman, pp. 161-7.

Garside, R. 1996. 'The robust tagging of unrestricted text: the BNC experience' in Thomas, J. and Short, M., eds. *Using corpora for language research: studies in the honour of Geoffrey Leech* Harlow: Longman, pp. 167-180.

Garside, R., Leech, G., and McEnery, T. (1997) 'Corpus Annotation: Linguistic Information from Computer Text Corpora' London: Longman, chapters 7-9.

Ide, N., Priest-Dorman, G., Véronis, J. (1996) *Corpus Encoding Standard*. [available from http://www.cs.vassar.edu/CES/]

Lee, D. (2001) 'Genres, registers, text types and styles: clarifying the concepts and navigating a path through the BNC Jungle' in *Language Learning and Technology*, 5.3. [available from http://llt.msu.edu/]

Leech, G., Garside, R., and Bryant, M. (1994). 'CLAWS4: The tagging of the British National Corpus' in *Proceedings of the 15th International Conference on Computational Linguistics (COLING 94)*. Japan: Kyoto, pp. 622-628.

Leech, G and Smith, N. (2000) *Manual to accompany The British National Corpus (Version 2) with Improved Word-class Tagging*. Lancaster: UCREL. [Supplied in digital form as part of the BNC World Edition].

Lehmann, H., Schneider, P., Hoffmann, S. (1999) 'BNCweb' in Kirk, J. ed. *Corpora galore: analysis and techniques in describing English*. Amsterdam: Rodopi, pp. 259-266.

Renouf, A. (1986) 'Corpus development at Birmingham University' in Aarts, J., and Meijs, W., eds. *Corpus Linguistics II: New Studies in the Analysis and Exploitation of Computer Corpora*, Amsterdam: Rodopi, pp. 7-23

Sinclair, J. McH. (1987) *Looking Up*. London: Collins.

Sperberg-McQueen, C.M., and Burnard, L. (1994) *Guidelines for electronic text encoding and interchange (TEI P3). Chicago and Oxford: ACH-ALLC-ACL Text Encoding Initiative.*

Zampolli, A., Calzolari, N., Palmer, M. (1994) eds. 'Current Issues in Computational Linguistics: In Honour of Don Walker', *(Linguistica Computazionale IX-X)*. Pisa: Giardini.