NOTES TO ACCOMPANY
THE BNC VERSION 1 (BIBLIOGRAPHICAL) INDEX

David Lee

Please Note: This manual was revised on 25 Oct 2001 so that all references to the BNC Index now read "BNC1 Index", to make it clear that this particular Index and all associated tables, word totals, etc. refer to version 1 (the first release) of the British National Corpus, rather than the more recent **BNC World Edition**, for which a separate spreadsheet and separate document similar to this one are available.

The BNC1 Index contains the most accurate and up-to-date information on the BNC files, and thus some information may supersede and/or be different from that contained in the official BNC files/file headers themselves. Nevertheless, for research purposes, users are strongly advised to use the BNCW Index and the BNC World Edition instead, as the newer version of the corpus contains fewer text classification and tagging errors.

The BNC1 Index spreadsheet was created as one solution to the problem of BNC 'domain' categories being overly broad and too inexplicit, to fix classification errors and steer people away from misleading file titles, and to provide a proper navigational map for people wanting to deal with specific 'genres' (as generally understood by most people). It is similar to the plain text ones prepared by Adam Kilgarriff which I have benefited from, and found rather useful. However, those files do not contain all the details which are needed for compiling your own sub-corpus (author type, author age, author sex, audience type, audience sex, section of text sampled, (topic) keywords, etc.)[1]. Sebastian Hoffmann's files[2] were useful too, in a complementary way, but these do not include (i) keywords, and (ii) the full bibliographical details of files. A third existing resource, the 'bncfinder.dat' file which comes with the standard distribution of the BNC (version 1) has most of the header information, but in the form of highly abbreviated numeric codes, and also does not include any bibliographical information about the files or keywords. The BNC1 Index consolidates the kinds of information available in the above three resources, but, in addition, includes: (i) BNC-supplied keywords (as entered in the file headers by the compilers); (ii) COPAC keywords[3] for

---

1 Kilgarriff's list only includes the first 80 characters or so of the title of each file, which means some titles are truncated (thus no good for searching by), and author names (for the written texts) are not included.

2 Available at ftp://escorp.unizh.ch/pub/bncstuff/databases/

3 COPAC is an on-line system for unified access to the (combined) catalogues of some of the largest university research libraries in the UK and Ireland (http://www.copac.ac.uk/). Keywords were manually copied from the web catalogue entries and put into a separate column in the BNC1 Index to allow researchers to search by proper library keywords in addition to the keywords provided by the BNC compilers. These keywords will greatly facilitate the identification of sub-genres, (sub-)topics, etc. by people who wish to have finer sub-classifications for specific research purposes.

published non-fiction texts[4] (topic keywords entered by librarians); (iii) full bibliographical details (including title, date and publisher for written texts, and number of participants for spoken files); (iv) an extra level of text categorisation, 'genre', where each text is assigned to one of the 70 genres or sub-genres (24 spoken and 46 written) developed for the purposes of this Index; (v) a column supplying 'Notes & Alternative Genres', where texts which are interdisciplinary in subject matter or which can be classified under more than one genre are given alternative classifications. Also entered here are extra notes about the contents of files (e.g. where a single BNC file contains several sub-genres within it, such as postcards, letters, faxes, etc., these are noted[5]). For some written texts taken from books, the title of the book series is also given under this column (e.g. file BNW, "Problems of unemployment and inflation", is part of the Longman book series "Key issues in economics and business").

It is hoped that this will be a comprehensive, user-friendly, 'one-stop' database of information on the BNC. All the information is presented using a minimum of abbreviations or numeric codes, for ease of use. For example, 'm_pub' (for 'miscellaneous published') is used instead of a cryptic number for the *medium* of the text, and *domains* are likewise indicated by abbreviated strings (e.g. 'W_soc_science', 'S_Demog_AB') rather than numbers. It should be noted that I carried out the genre categorisation of all the texts by myself: this ensures consistency, but it also means that some decisions may be debatable. The pragmatic point of view I am taking is that something is better than nothing, and that it is beneficial to start with a reasonable genre categorisation scheme and then let end-users report problem/errors and dictate future updates and improvements.

When compiling a sub-corpus for the purpose of research, classroom concordancing, genre-based learning, etc., you need all the available information you can get. With the BNC1 Index, it is now possible, for example, to separate *children's prose fiction* from *adult prose fiction* by combining information from the 'audience age' field and the newly introduced 'genre' field (using *domain* alone would have included poems as well).

All the information in the spreadsheet is up-to-date and as accurate as possible, and supersedes the information given in the actual file headers and the 'bncfinder.dat' file distributed with the BNC (version 1), both of which are known to contain many errors[6]. Changes and corrections to erroneous classifications were made after extensive manual checks, and on the basis of error reports made by others. The following section lists and explains all the columns/fields of information given in the BNC1 Index.

---

4 For an explanation of why only non-fiction works are given keywords, see the British Library 'Fiction Indexing Policy' document at http://www.bl.uk/services/bsds/nbs/marc/655polc.html.

5 These extra notes are the result of random, manual checks: not all files have been subjected to such detailed analysis.

6 BNC World Edition headers should have been updated and purged of these errors.

*Notes on the **BNC1 Index***

For **spoken** files, there are only 8 relevant fields of information, giving the following self-explanatory details (abbreviations are expanded below in Table 1):[7]

File ID	*Domain*	*Genre*	*Keywords*	*Word Total*	*Interaction Type*	*Mode*	*Bibliographical Details*
FLX	S_cg_education	S_classroom	natural & pure science; chemistry	5142	Dialogue	S	11th year science lesson: lecture in chemistry of metal processing (Edu/inf). Rec. on 23 Mar 1993 with 2 partics, 381 utts

Note that *Mode* only distinguishes broadly between spoken (S) and written (W). To further restrict searches to only 'demographic' files or only 'context-governed' files, the *Domain* field should be used.

For **written** files, there can be up to 19 fields of information (depending on the file: fields which do not apply to a particular file are left blank). As an example, the entry for AE7 is:

File ID	*Medium*	*Domain*	*Genre*	*Notes & Alter-native Genres*	*COPAC Keywords*	*Keywords*	*Audience Age*	*Audience Sex*	*Audience Level*	*Bibliographical details*
AE7	book	W_nat_science	W_non_ac_nat_science	Also W_non_ac_humanities_arts	Biology - Philosophy	molecular genetics	adult	mixed	high	The problems of biology. Maynard Smith, John. Oxford: OUP, 1989, pp. 9-109. 1686 s-units.

(Table: continued)

Total Words	*Sampling*	*Circulation Status*	*Time Period*	*Mode*	*Author Age*	*Author Sex*	*Author Type*
36115	mid	M	1985-1994	W	60+ yrs	Male	Sole

Some of the information fields are explained more fully in the *BNC User's Reference Guide*, but here is a brief explanation:

The table above tells us that file **AE7** is a sample extracted from the **mid**dle (*Sample Type*) of a **book** (*Medium*), whose *Circulation Status* is **M**edium (this refers to the number of receivers of the text)[8], whose author (*Author*

---

7 Note that for the demographic files (conversations) the *Keywords* field is empty for almost all the files.

8 The somewhat confusing term *reception status* is used in the *BNC Users' Reference Guide* instead of *circulation status*. Since it refers to the **size** of the readership or the circulation level (not the social status of the text), I have changed the label to reflect this. Circulation status should be used with caution, because it is **relative** to genre: a newspaper with 'low' reception status may still have a lot more readers than a 'medium-reception' book of poetry or office memo. The field *(Target) Audience level*, on the other hand, is an **estimate** (by the compilers) of the **level of difficulty** of the text, or the amount of background knowledge of its subject matter which is assumed.

Age/Sex/Type) is **60+ yrs** old (age band 6 in terms of BNC codes), is **Male** and is the **Sole** author of the text. The text has been manually classified as '**non_ac**ademic prose, **nat**ural **sc**iences' (*Genre*), although it also deals with philosophical issues (*COPAC Keywords*) and thus may also be considered under 'W_non_ac_humanities_arts'. The target audience for the text are **adults**, of both sexes (**mixed**), and **high**-level (original BNC numerical code='level 3'). The BNC compilers have classified it under '**nat**ural **sc**iences' (*Domain*)[9], and the text was composed in the period 1985-1994 (*Time Period*)[10]. The *Bibliographical Details* field gives us the title of the text ('The Problems of Biology'), its author, publisher, etc., and an indication of the number of sentences ('s-units'), while the [BNC compilers'] *Keywords* field supplies the detail that the book is about molecular genetics (COPAC and BNC keywords tend to be about **topic**, and are sometimes useful for **sub-genre** identification). The page numbers under Bibliographical Details were, in this case and many others, not actually given in the original BNC bibliography, but were manually added to the Index after I had searched in the file for the page-break SGML elements. This is to allow proper, complete referencing[11].

A list of all possible values for the closed-set fields (the keyword fields are open-ended) is given below (the abbreviations/codes are in bold):

Field	**Possible Values**
Medium	[Written texts only] **book, m_pub** (miscellaneous, published), **m_unpub** (miscellaneous unpublished), **periodical** (magazines, journals, etc.), **to_be_spoken** (written-to-be-spoken)
Domain	**S_cg_business** (context-governed, business), **S_cg_education** (c-g, educational), **S_cg_leisure** (c-g, leisure), **S_cg_public** (c-g, public/institutional), **S_Dem_AB/C1/C2/DE/Unc** (spoken demographic classes for the casual conversation files; 'Unc' = 'unclassified'), **W_app_science** (applied science), **W_arts**, **W_belief_thought** (belief & thought), **W_commerce** (commerce & finance), **W_imaginative** (imaginative/creative), **W_leisure** (leisure), **W_nat_science** (natural sciences), **W_soc_science** (social sciences), **W_world_affairs** (world affairs).
Genre (70 in total)	*[Spoken texts, 24 genres]:* **S_brdcast_discussn** (TV or radio discussions), **S_ brdcast_documentary** (TV documentaries), **S_brdcast_news** (TV or radio news broadcasts), **S_classroom** (non-tertiary classroom discourse), **S_ consult** (*mainly* medical & legal

9 Note that *Genre* classifications (assigned by me) do not always agree with the *Domain* classifications of the BNC compilers (i.e. the official domain classifications as given in the standard distribution of the corpus).

10 This follows the new 4-way time period classification employed in the BNC World Edition: alltim0 (**---[unclassified]**); alltim1 (**1960-1974**); alltim2 (**1975-1984**); alltim3 (**1985-1994**). The old classification code 'writim' was for written texts only.

11 The original bibliographical references were 'pp. ??'. Some files did not have page breaks encoded at all, however, and thus their bibliographical references remain incomplete.

consultations), **S_conv** (face-to-face spontaneous conversations), **S_courtroom** (legal presentations or debates), **S_demonstratn** ('live' demonstrations), **S_interview** (job interviews & other types), **S_interview_oral_history** (oral history interviews/narratives, some broadcast), **S_lect_commerce** (lectures on economics, commerce & finance), **S_lect_humanities_arts** (lectures on humanities and arts subjects), **S_lect_ nat_science** (lectures on the natural sciences), **S_lect_polit_law_edu** (lectures on politics, law or education), **S_lect_soc_ science** (lectures on the social & behavioural sciences), **S_meeting** (business or committee meetings), **S_parliament** (BNC-transcribed parliamentary speeches), **S_pub_debate** (public debates, discussions, meetings), **S_sermon** (religious sermons), **S_speech_scripted** (planned speeches), **S_speech_unscripted** (more or less unprepared speeches), **S_sportslive** ('live' sports commentaries and discussions), **S_tutorial** (university-level tutorials), **S_unclassified** (miscellaneous spoken genres).

[Written texts, 46 genres]

W_ac_humanities_arts (academic prose: humanities), **W_ac_medicine** (academic prose: medicine), **W_ac_nat_ science** (academic prose: natural sciences), **W_ac_polit_law_edu** (academic prose: politics, law, education), **W_ac_soc_ science** (academic prose: social & behavioural sciences), **W_ac_tech_engin** (academic prose: technology, computing, engineering), **W_admin** (adminstrative and regulatory texts, in-house use), **W_advert** (print advertisements), **W_biography** (biographies/autobiographies), **W_commerce** (commerce & finance, economics), **W_email** (e-mail sports discussion list), **W_essay_school** (school essays), **W_essay_univ** (university essays), **W_fict_drama** (excerpts from two modern drama scripts)**, W_fict_poetry** (single- and multiple-author collections of poems)**, W_fict_prose** (novels & short stories), **W_hansard** (Hansard/parliamentary proceedings), **W_institut_doc** (official/govermental documents/leaflets, company annual reports, etc.; excludes Hansard), **W_instructional** (instructional texts/DIY), **W_letters_personal** (personal letters, postcards, notes)**, W_letters_prof** (professional/business letters), **W_misc** (miscellaneous texts), **W_news_script** (TV autocue data), **W_newsp_brdsht_nat_arts** (broadsheet national newspapers: arts/cultural material), **W_newsp_brdsht_nat_commerce** (broadsheet national newspapers: commerce & finance), **W_newsp_brdsht_nat_editorial** (broadsheet national newspapers: personal & institutional editorials, & letters-to-the-editor), **W_newsp_brdsht_nat_misc** (broadsheet national newspapers: miscellaneous material), **W_newsp_brdsht_nat_report** (broadsheet national newspapers: home & foreign news reportage), **W_newsp_brdsht_nat_science** (broadsheet national newspapers: science material), **W_newsp_brdsht_nat_social** (broadsheet national newspapers:

	material on lifestyle, leisure, belief & thought), **W_newsp_brdsht_nat_sports** (broadsheet national newspapers: sports material), **W_newsp_other_arts** (regional and local newspapers: arts), **W_newsp_other_commerce (**regional and local newspapers: commerce & finance)**,** **W_newsp_other_report** (regional and local newspapers: home & foreign news reportage)**, W_newsp_other_science (**regional and local newspapers: science material)**, W_newsp_other_social (**regional and local newspapers: material on lifestyle, leisure, belief & thought)**, W_newsp_other_sports, W_newsp_tabloid** (tabloid newspapers)**, W_non_ac_humanities_arts** (non-academic/non-fiction: humanities), **W_non_ac_medicine** (non-academic: medical/health matters), **W_non_ac_nat_science** (non-academic: natural sciences), **W_non_ac_polit_law_edu** (non-academic: politics, law, education), **W_non_ac_soc_ science** (non-academic: social & behavioural sciences), **W_non_ac_tech_engin** (non-academic: technology, computing, engineering), **W_pop_lore** (popular magazines), **W_religion** (religious texts, excluding philosophy).
Mode	**W** (written), **S** (spoken)
Author age	**0-14 yrs** (band 1), **15-24 yrs** (band 2), **25-34 yrs** (band 3), **35-44 yrs** (band 4), **45-59 yrs** (band 5), **60+ yrs** (band 6), **---** (unclassified)
Author sex	**Male**, **Female**, **Mixed**, **Unknown**, **---** (not applicable/available)
Author type	**Corporate**, **Multiple**, **Sole**, **Unknown**/unclassified
Audience age	**child**, **teen**, **adult**, **---** (unclassified)
Audience sex	**male**, **female**, **mixed**, **---** (unclassified)
Audience level	**low** (level 1), **medium** (level 2), **high** (level 3), --- (unclassified)
Sampling	whole text (**whl**), beginning sample (**beg**), middle sample (**mid**), end sample (**end**), composite (**cmp**), unknown/not applicable (**--**).
Circulation Status	(formerly 'reception status'): **L**ow, **M**edium, **H**igh (blank for unclassified texts)

Table 1 Information fields and possible values in the BNC1 Index

With all these fields of information put together in a one database/spreadsheet, where they can be combined with one another, it becomes easy to scan the BNC for whatever particular kinds of text you are interested in.

Further notes on the genre classifications

The genre categories used in the BNC1 Index were chosen after a survey of the genre categorisation schemes of other existing corpora (e.g. LLC, LOB, ICE-GB) and will thus be familiar to users and compatible with these other corpora, allowing comparative studies based on genres taken from different corpora. These genre labels have been carefully selected to capture as wide a range as possible of the numerous types of spoken and written texts in the English language, and the divisions are more fine-grained than the domain categories

used in the BNC itself. Note that some genre labels are **hierarchically nested**, so that, for example, if you simply want to study 'prototypical academic English' and are not concerned with the sub-divisions into social sciences, humanities, etc., you can find all such files by searching for 'W_ac*' and specifying 'high' for 'audience level'[12]. Or if you are interested in the language of the social sciences, whether spoken or written, you can similarly use wildcards to search for '*_soc_science'. In general, where further sub-genres can be generated on-the-fly through the use of other classificatory fields, they are not given their own separate genre labels, to avoid clutter. For instance, 'academic texts' can be further sub-divided into '(introductory) textbooks' and 'journal articles', but since this can be done easily by using the *medium* field (viz. choosing either 'book' or 'periodical'), the sub-genres have not been given their own separate labels. Instead, end-users are encouraged to use all available fields to create their own sub-classificatory permutations. As another example, the broad genre 'institutional documents' includes government publications (including 'low-brow' informational booklets and leaflets/brochures), company annual reports, and university calendars and prospectuses. These texts can be fairly easily separated out using 'Medium', 'Audience level' or 'Keywords', however. The fuzzy 'genre' labels are therefore meant to provide starting points, not a definitive taxonomy.

The tables below show the breakdown of the genre categories used in the BNC1 Index spreadsheet more clearly than in the earlier table, and also shows the 'super-genres' that some researchers may want to study (made possible by the use of hierarchically nested genre labels).

BNC1 SPOKEN	No. of words	%	Big Genre	# of Files
S_brdcast_discussn	761,187	7.4%		54
S_brdcast_documentary	41,509	0.4%	Broadcast	10
S_brdcast_news	261,039	2.5%	10.3%	12
S_classroom	435,370	4.2%		59
S_consult	137,821	1.3%		128
S_conv	4,211,216	40.7%		153
S_courtroom	127,331	1.2%		13
S_demonstratn	31,736	0.3%		6
S_interview	123,679	1.2%	Interviews	13
S_interview_oral_history	814,283	7.9%	9.1%	119
S_lect_commerce	15,105	0.1%		3
S_lect_humanities_arts	50,762	0.5%		4
S_lect_nat_science	22,650	0.2%	Lectures	4
S_lect_polit_law_edu	50,849	0.5%	2.8%	7
S_lect_soc_science	159,689	1.5%		13
S_meeting	1,376,072	13.3%		132
S_parliament	96,111	0.9%		6
S_pub_debate	283,231	2.7%		16
S_sermon	82,185	0.8%		16
S_speech_scripted	200,072	1.9%	Speeches	26

12 The use of this additional specification, 'audience level=high', will roughly filter out introductory textbooks and texts written for mixed or more general audiences.

S_speech_unscripted	464,287	4.5%	6.4%	51
S_sportslive	33,060	0.3%		4
S_tutorial	143,084	1.4%		18
S_unclassified	421,148	4.1%		44
TOTAL	10,343,476	*100.00%*		*911*

Table 2 Breakdown of Spoken BNC Version 1 Genres

N.B. The following 4 spoken files were not genre-categorised in the BNC1 Index because they contain duplicated/repeated material: D98, HDE, HDF, HDG. Hence there are 911 genre-categorised spoken files instead of 915. (In BNC World, all except HDE were removed.)

BNC1 WRITTEN	**No. of words**	**%**	**Big Genre**	**# of Files**
W_ac_humanities_arts	3,319,624	3.7%		87
W_ac_medicine	1,421,802	1.6%	Academic	24
W_ac_nat_science	1,111,311	1.2%	Prose	43
W_ac_polit_law_edu	4,677,764	5.2%	17.4%	187
W_ac_soc_science	4,406,825	4.9%		142
W_ac_tech_engin	685,613	0.8%		23
W_admin	219,844	0.2%		12
W_advert	557,896	0.6%		60
W_biography	3,524,002	3.9%		100
W_commerce	3,757,766	4.2%		112
W_email	212,999	0.2%		7
W_essay_sch	146,474	0.2%	Non-printed	7
W_essay_univ	65,385	0.1%	essays	4
W_fict_drama	45,735	0.1%		2
W_fict_poetry	230,244	0.3%	Fiction	31
W_fict_prose	18,021,871	20.1%	20.5%	485
W_hansard	1,155,709	1.3%		4
W_institut_doc	546,039	0.6%		43
W_instructional	436,585	0.5%		15
W_letters_personal	52,422	0.1%	Letters	6
W_letters_prof	65,993	0.1%	0.2%	11
W_misc	9,161,238	10.2%		501
W_news_script	1,294,044	1.4%		32
W_newsp_brdsht_nat_arts	351,533	0.4%		51
W_ newsp_brdsht_nat _commerce	424,822	0.5%		44
W_ newsp_brdsht_nat _editorial	101,626	0.1%		12
W_ newsp_brdsht_nat _misc	1,032,462	1.2%	Broadsheet	95
W_ newsp_brdsht_nat _reportage	663,133	0.7%	National	49
W_ newsp_brdsht_nat _science	65,253	0.1%	Newspapers	29
W_ newsp_brdsht_nat _social	81,820	0.1%	3.4%	36
W_ newsp_brdsht_nat _sports	297,575	0.3%		24
W_newsp_other_arts	239,054	0.3%		15
W_newsp_other_commerce	415,237	0.5%	Regional &	17
W_newsp_other_report	2,716,252	3.0%	Local	39
W_newsp_other_science	54,789	0.1%	Newspapers	23

W_newsp_other_social	1,142,371	1.3%	6.3%	37
W_newsp_other_sports	1,027,219	1.1%		9
W_newsp_tabloid	728,234	0.8%	Tabloids	6
W_non_ac_ humanities_arts	3,924,117	4.4%		116
W_non_ac_medicine	498,157	0.6%		17
W_non_ac_nat_science	2,506,378	2.8%	Non-academic	62
W_non_ac_polit_law_edu	4,475,898	5.0%	Prose	93
W_non_ac_soc_science	4,176,025	4.7%	18.8%	128
W_non_ac_tech_engin	1,209,661	1.3%		123
W_pop_lore	7,371,119	8.2%		211
W_religion	1,120,624	1.2%		35
TOTAL	89,740,544	100.0%		3209

Table 3 Breakdown of Written BNC Version 1 Genres

It will be noted that aspects of this genre classification scheme mirror the ICE-GB corpus, although I have made finer distinctions in some cases (e.g. the lecture and broadsheet sub-genres) and grouped texts differently (e.g. I have 'nested' all the broadsheet newspaper material together rather than split it into separate functional groups as in the ICE-GB (which has 'Reportage' and 'Persuasive writing').

In some respects, the scheme also follows the Lancaster-Oslo/Bergen (LOB) corpus quite closely. This was done deliberately, to facilitate diachronic/comparative research.[13] For example, here is how the various subject disciplines are categorised in the LOB corpus and in the BNC1 Index:

LOB (& BNC1 Index) Category	**Subjects/Disciplines Represented**
Humanities	Philosophy, History, Literature, Art, Music
Social sciences	Psychology, Sociology, Linguistics, Social Work
Natural sciences	Physics, Chemistry, Biology
Medicine	-
Politics, Law, & Education	-
Technology & Engineering	Computing, Engineering

Table 4 LOB corpus categories broken down into component disciplines

One difference from the LOB corpus is that **economics** texts in the BNC1 Index are not put under 'politics, law and education', but are instead put under the 'W_commerce' genre. Also, **archaeology** and **architecture** have been classified as humanities or arts subjects under the present scheme, **geology** has been classed as a natural science, and **geography** is classed either as a social or natural science depending on the branch of geography involved. One **mathematics** textbook file for primary/elementary schools was simply put under 'miscellaneous', while university-level mathematical texts were put under either 'natural_sciences' or

13 The LOB corpus already has, of course, a modern-day correlative: the FLOB (Freiburg LOB) corpus. My categorisations will allow the BNC to also be used in comparative studies alongside these corpora.

'technology & engineering' depending on whether they were pure or applied.[14] It should also be noted that some texts represent a mixture of disciplines (e.g. history and politics often go hand in hand, but the two are separate categories under this scheme). In such cases, a more or less arbitrary assignment was made, based on what was judged to be the dominant point of view in the text, and, in the case of printed publications, after consultation of the keywords for the text in library catalogues.

The 'non-academic' genres relate to written texts (mainly books) sometimes called 'non-fiction' which have subject matters belonging to one of the disciplines listed above. They are usually texts written for a general audience, or 'popularisations' of academic material, and are thus distinguished from texts in the parallel academic genres (which are targeted at university-level audiences, insofar as this can be determined). In deciding whether a text was academic or not, a variety of cues was used: (i) the 'audience level (of difficulty)' estimated by the BNC compilers (coded in the file headers) (ii) whether COPAC lists the book as being in the 'short loan' collections of British universities (this works in one direction only: absence is not indicative of a work not being academic) (iii) the publisher and publication series (academic publishers form a small and recognisable set, and some books have academic series titles, which help to place them in context).

Deciding what a coherent genre or sub-genre is can be far from easy in practice, as (sub-)genres can be endlessly multiplied or sub-divided quite easily. Moreover, the classificatory decisions of corpus compilers may not necessarily be congruent with that of researchers. For example, what is considered 'applied science'? In the present scheme, 'applied science' excludes medicine (which is instead placed in a category of its own), engineering (which is put under 'technology'), and computer science (also under 'technology'). For the purposes of the BNC1 Index, a particular 'level of delicacy' has been decided on for the genre scheme, based on categories already in use in existing corpora and in the research literature. Users may further sub-divide or collapse/combine genres as they see fit. The present scheme is only an aid, to help narrow down the scope of any sub-corpus building task. In this connection, it should be noted that due to the way the material was recorded and collated, many of the spoken **files** (especially 'conversation' ones) are less well-defined than the written files because they are made up of different task and goal types, as well as varying topics and participants (e.g. a single 'conversation' file can contain casual talk between both equals and unequals, and 'lecture' files often contain casual preambles and concluding remarks in addition to the actual lectures themselves). Researchers wanting discoursally well-defined and homogeneous texts will have to sub-divide **files** into **texts** themselves.

If the distribution of linguistic features among 'genres' is important to a particular piece of research, then obviously the research can be affected or compromised by the definition/constitution of the 'genres' in the first place. For this reason, users of the BNC1 Index are advised to read the notes/documentation given here, and to be clear what the various domain and genre labels mean.[15] To

---

14 People who disagree with these classifications may use the 'Keywords' and 'Title' fields to find the relevant files and re-classify them as desired.

15 The *domain* labels in the BNC1 Index are largely unchanged (i.e. they reflect the decisions of the BNC compilers). Some egregious errors were corrected, however, and

illustrate: the BNC compilers have classified some texts into the 'natural/pure sciences' domain (e.g. text CNA, which is taken from the *British Medical Journal*) which I would consider as belonging to 'applied science' or else simply 'medicine' as a separate category. On the other hand, the BNC compilers appear to have a rather loose definition of what 'applied science' is. Anything which is not directly classifiable or recognisable as being purely about theoretical physics, chemistry, biology or medicine is apparently considered 'applied'. For example, consider:

Text ID	***Medium***	***Domain***	***Bibliographical Details***
FYX	book	W_app_science	Black holes and baby universes. Hawking, Stephen W. London: Bantam (Corgi), 1993, pp. 1-139. 1927 s-units.
AMS	book	W_app_science	Global ecology. Tudge, Colin. London: Natural History Museum Pub, 1991, pp. 1-98. 1816 s-units.
AC9	book	W_app_science	Science and the past. London: British Museum Press, 1991, pp. ??. 1696 s-units.

The first book is a popularisation by Stephen Hawking and is an application of physics to the study of the universe or outer space. In the BNC1 Index genre scheme, I would consider this to be part of the 'non-academic **natural sciences**' genre (rather than 'applied science'). It is a similar situation with the second and third books (which concern ecology and archaeological/historical work respectively). It is true that these are also about applying scientific ideas in some way, but they do not quite fit in with the more common understanding of 'applied science'. In the present scheme, text AMS would be under 'academic: natural science', and AC9 under 'non-academic: humanities'.

As another example of the classificatory system used here, consider the case of linguistics. Some linguists, including myself, would consider our discipline to be a social science (although others would place us in the humanities). In any case, consider the way the following BNC texts were (inconsistently) classified by the compilers:

Text ID	*Medium*	***Domain***	*Details*
B2X	periodical	W_app_science	Journal of semantics. Oxford: OUP, 1990, pp. 321-452. 847 s-units.
CGF	book	W_arts	Feminism and linguistic theory. Cameron, Deborah. Basingstoke: Macmillan Pubs Ltd, 1992, pp. 36-128. 1581 s-units.
EES	m_unpub	W_app_science	Large vocabulary semantic analysis for text recognition. Rose, Tony Gerard. u.p., n.d., pp. ??. 2109 s-units.
FAC	book	W_soc_science	Lexical semantics. Cruse, D A. Cambridge: CUP, 1991, pp. 1-124. 2261 s-units.
FAD	book	W_soc_science	Linguistic variation and change. Milroy, J. Oxford: Blackwell, 1992, pp. 48-160. 1339 s-units.

It may be the case that the actual content/topic of these linguistics-related texts make them seem less like social science texts than arts or applied science texts (e.g. text ESS is a dissertation on computer handwriting recognition by a student from a department of computing,). But if so, what does it make of the general public's understanding of domain labels like 'linguistics' and 'social sciences', then? These are important questions when one is seeking to draw conclusions

reported to the BNC project for fixing in the new release, BNC World Edition.

about the distribution of linguistic features found in particular genres. For the present purposes, therefore, one particular stand has been taken on how to classify texts, and readers should bear this in mind. (In the case of the above example, all were classified as 'academic: social science' except EES, which was put under 'academic: technology and engineering'.)

*Using the **BNC1 Index***

The BNC1 Index will be distributed in the Microsoft Excel® spreadsheet format as well as in a tab-delimited format (it will also be incorporated into two custom-built, user-friendly programs: see below).[16] On a practical note, the advantage of using the Excel format is that there is a quick way of displaying only the texts which match your chosen criteria through the use of the relatively user-friendly 'Autofilter' function (under the 'Data' menu in the program, choose 'Filter' and then 'Autofilter'). With the Autofilter switched on, the top row of every field (column) will have a drop-list which can be used to instantly filter down to the texts you want displayed (clicking on the drop-list button reveals all the possible values for that field (e.g. genre), and you just select the one you want). Fields are combinable, so you can, for example, first restrict the display to only 'social science' texts under *domain*, then further restrict this to only 'periodicals' under *medium,* and end up with social science periodicals. It is also possible to make more **advanced** searches, by activating the *'*Custom*'* filter dialogue box from the relevant drop-list. This will allow you to filter the fields using wildcards. One caveat needs to be issued to users, however: they should not rely entirely on the genre labels, but should also check the 'Alternative Notes' column and scan/browse the files too. For example, some files labelled "S_brdcast_discussion" actually also contain *news reports*, because these are sometimes broadcast in the middle of a radio broadcast discussion or talk show. This serves as a warning that some BNC files combine genres and sub-genres but can only be labelled in terms of the *majority* type. Some of the BNC-supplied fields are also not entirely accurate. Many of the files coded as 'monologue' (under the *Interaction Type* column), for example, actually include some dialogue as well (i.e. they are **mostly** monologue, but not exclusively).

A stand-alone Windows® program, called *BNC Indexer*®, **for the BNC Version 1**, has been developed by Antonio Moreno Ortiz using the information contained in my spreadsheet.[17] A web-based facility, *BNC Web Indexer* (for **BNC Version 1 only**, at the time of writing), has also been developed at Lancaster which does essentially the same thing[18]. Both programs are similar in layout and function. They are much easier to use than the Excel spreadsheet since they do not require any knowledge of spreadsheet/database programs and have very simple, intuitive interfaces (perfect for classroom situations). All the information fields (*domain, genre, audience age, author sex*, etc.) and their values are displayed on screen

---

16 The BNC1 Index spreadsheet should be referenced as being available from: http://clix.to/davidlee00

17 Available at: http://webdeptos.uma.es/filifa/personal/amoreno/indexer/ It is freeware. **Note, however, that it is not updated as often as the BNC Index itself, and is already out of date.**

18 *BNC Web Indexer* is the result of a collaboration between Paul Rayson (UCREL, Lancaster University) and myself.

and users simply select the values they want to use and then press a button to execute the query. A results panel shows all the texts which match the filtering criteria, along with bibliographical and other information. (With *BNC Indexer* (for BNC version 1), individual texts can also be deselected from the output list if so desired, and can be browsed first by double-clicking on the relevant line.) Output file lists containing the file IDs of the BNC files which matched the criteria can be generated and fed into concordancers such as WordSmith or MonoConc[19] which can use a list of filenames to specify a sub-corpus to which future queries are to be restricted. Note that with both *BNC Indexer* and *BNC Web Indexer*, individual files can always be deleted from the output list if so desired, so users do not have to accept the classification decisions wholesale but can vet individual texts before allowing them into a sub-corpus.

Known BNC Errors

1) In BNC Version 1, files **G3M** and **G3J** contained the wrong text bodies (i.e. the file headers and bibliographical details were correct, but the texts were swapped around). Thus, **G3M** contained the text of the novel "Heat and Dust", while G3J contained "The Licensing (Scotland) Act 1976". In BNC World Edition, however, both files contain the same text, viz. the Licensing Act text (the fiction text appears to have vanished).

2) The following 4 spoken files are not genre-categorised in the BNC1 Index because they contain duplicates/repeated material: **D98, HDE, HDF**, **HDG**. Hence there are 911 genre-categorised spoken files instead of 915. (In BNC World, all except **HDE** were removed.).

3) In **both** BNC Version 1 and BNC World Edition, file **HEP** (which contains material from **HDE, HDF** and **HDG**) has the header description "Oral history project: interview", but this is obviously wrong. In the both BNC1 Index and BNCW Index, HEP is instead given the title "Enterprise Two Thousand: seminar" and genre-classified under S_speech_unscripted. This, along with many other changes, may not be reflected in the actual file headers of the BNC texts.

4) [4 May 2002] File **JP8**: This was erroneously re-classified as 'domain=S_cg_business' in my previous edition of BNC World Index (but not the BNC1 Index). In fact, the official BNC classification, 'S_cg_education' is the correct one. I've now fixed this. I've also changed the Interaction Type to 'Dialogue', as this appears to be classroom/tutorial interactive discourse, and changed the file title from "Training session. Rec. on 13 Jan 1994 with 2 partics, 166 utts" to "Classroom lesson. Sample containing about 2721 words speech recorded in educational context.".

5) [4 May 2002] I am adding a note here to indicate that file **JSE** has the wrong domain classification ("spoken context-governed, educational/informative") in the official BNC (both versions). In my Index, it is instead classified as "spoken context-governed, business", since this is a supermarket training seminar/course for employees.

---

19 Or using the web-based concordancer for the BNC developed at Zürich, *BNCweb*, at http://escorp.unizh.ch (restricted usage). The new version of SARA developed for the BNC World Edition is also expected to have more sophisticated sub-corpus querying facilities.

6) [10 May 2002] While changing the title and genre category for file **HEP** (see note 3 above), I neglected to change the domain category from 'S_cg_leisure' to 'S_cg_business'. This has now been rectified.

[The following are errors reported by Sebastian Hoffman to the BNC-discuss list on 6th June 2001.]

7) In file AK4, sentences 1025-1129 are an exact repetition of sentences 918-1023.

8) CCP is almost exactly the same as CJV - there are just a few minor mark-up differences and 12 words less in CJV (according to the header, which otherwise has the same information for both texts). *[I think the markup is better in file CCP. – DL]*

9) Sentences 577-632 in file CRP are repeated as sentences 734-790 in FT7.

10) Files HJH and H8N are pretty much the same. The header differs in one peculiar way: For HJH, the publication date is given as 1985-1993 and for H8N as 1975-1984. It seems that sentences 1-3580 in HJH are pretty much the same as sentences 1-3573 in H8N. HJH then adds about 300 sentences entitled "Part three - five years later".

11) In file HH3, sentences 11354-11365 are repeated as sentences 11635- 11647.

12) In file K1Y, sentences 187-278 are repeated as sentences 279 - 370.

13) Sentences 3148-3174 in file K2D are repeated as sentences 3345-3371 in K32.

14) K27: sentences 779 - 832 are repeated as 960-1026 with slightly different headings but otherwise apparently identical content.

15) KBK: sentences 5602-5608 are repeated in the same file as sentences 7008-7014. But the speakers are different. First it's David and Chris (2 male speakers) and later it's Chris (who now takes the part of David in the previous dialogue) and Norrine (a woman).

~~ * ~~

Last edited: 10 May 2002 (updated web URLs, added caveat + other changes)

Previous edition: 25 Oct 2001

Send Feedback, Comments, Error Reports to: david_lee00@hotmail.com

An expanded version of this document may be found in *Language Learning & Technology* (Special issue on "Using Corpora in Language Teaching and Learning", September 2001, Vol. 5, No. 3). It discusses my use of the term 'genre', compares the BNC genre scheme used here with the text categorisation schemes used in other corpora, and discusses some problems with the BNC which the genre categorisation of texts reported here attempts to solve.