

*Last edited: 30 Oct 2002*

*Previous editions: 10 May 2002, 4 May 2002, 25 Oct 2001, 13 Aug 2002*

## *NOTES TO ACCOMPANY*

### *THE **BNC** WORLD EDITION (BIBLIOGRAPHICAL) INDEX*

---

*[David Lee](#)*

**Please Note:** In this manual, all references to the BNC Index use the form “[BNCW Index](#)”, to make it clear that this particular Index and all associated tables, word totals, etc. refer to the more recent BNC World Edition, rather than BNC version 1 (the first release) of the British National Corpus, for which a [separate spreadsheet](#) and separate document similar to this one are available. The newer version of the corpus contains fewer text classification and tagging errors.

The [BNCW Index](#) aims to contain 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**.

The [BNCW 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.)<sup>1</sup>. Sebastian Hoffmann’s files<sup>2</sup> 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 [BNCW 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<sup>3</sup> for published non-fiction

---

<sup>1</sup> 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.

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

<sup>3</sup> 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

texts<sup>4</sup> (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<sup>5</sup>). 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 [BNCW 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<sup>6</sup>. 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 [BNCW Index](#).

---

column in the [BNCW 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.

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 [BNCW 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):<sup>7</sup>

<i>File ID</i>	<i>Domain</i>	<i>Genre</i>	<i>Keywords</i>	<i>Word Total</i>	<i>Interaction Type</i>	<i>Mode</i>	<i>Bibliographical Details</i>
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 parts, 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:

<i>File ID</i>	<i>Medium</i>	<i>Domain</i>	<i>Genre</i>	<i>Notes &amp; Alternative Genres</i>	<i>COPAC Keywords</i>	<i>Keywords</i>	<i>Audience Age</i>	<i>Audience Sex</i>	<i>Audience Level</i>	<i>Bibliographical details</i>
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)

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

Some of the information fields are explained more fully in the [BNC Users’ Reference Guide](#), but here is a brief explanation:

The table above tells us that file **AE7** is a sample extracted from the **middle** (*Sample Type*) of a **book** (*Medium*), whose *Circulation Status* is **Medium** (this refers to the number of receivers of the text)<sup>8</sup>, whose author (*Author 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.

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

<sup>8</sup> 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.

The text has been manually classified as ‘**non\_academic** prose, **natural sciences**’ (*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 ‘**natural sciences**’ (*Domain*)<sup>9</sup>, and the text was composed in the period 1985-1994 (*Time Period*)<sup>10</sup>. 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<sup>11</sup>.

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] <b>book</b> , <b>m_pub</b> (miscellaneous, published), <b>m_unpub</b> (miscellaneous unpublished), <b>periodical</b> (magazines, journals, etc.), <b>to_be_spoken</b> (written-to-be-spoken)
Domain	<b>S_cg_business</b> (context-governed, business), <b>S_cg_education</b> (c-g, educational), <b>S_cg_leisure</b> (c-g, leisure), <b>S_cg_public</b> (c-g, public/institutional), <b>S_Dem_AB/C1/C2/DE/Unc</b> (spoken demographic classes for the casual conversation files; ‘Unc’ = ‘unclassified’), <b>W_app_science</b> (applied science), <b>W_arts</b> , <b>W_belief_thought</b> (belief & thought), <b>W_commerce</b> (commerce & finance), <b>W_imaginative</b> (imaginative/creative), <b>W_leisure</b> (leisure), <b>W_nat_science</b> (natural sciences), <b>W_soc_science</b> (social sciences), <b>W_world_affairs</b> (world affairs).
Genre (70 in total)	[ <i>Spoken texts, 24 genres</i> ]: <b>S_brdrctst_discussn</b> (TV or radio discussions), <b>S_brdrctst_documentary</b> (TV documentaries), <b>S_brdrctst_news</b> (TV or radio news broadcasts), <b>S_classroom</b> (non-tertiary classroom discourse), <b>S_consult</b> ( <i>mainly</i> medical & legal consultations), <b>S_conv</b> (face-to-face spontaneous conversations), <b>S_courtroom</b> (legal presentations or debates), <b>S_demonstratn</b> (‘live’ demonstrations), <b>S_interview</b> (job interviews & other types), <b>S_interview_oral_history</b> (oral history interviews/narratives, some

<sup>9</sup> 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).

<sup>10</sup> 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.

<sup>11</sup> The original bibliographical references were ‘pp. ??’. Some files did not have page breaks encoded at all, however, and thus their bibliographical references remain incomplete.

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 speech, whether dialogue or monologue), **S\_speech\_unscripted** (more or less unprepared speech, whether dialogue or monologue), **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** (administrative 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/governmental 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\_brdsh\_t\_nat\_arts** (broadsheet national newspapers: arts/cultural material), **W\_newsp\_brdsh\_t\_nat\_commerce** (broadsheet national newspapers: commerce & finance), **W\_newsp\_brdsh\_t\_nat\_editorial** (broadsheet national newspapers: personal & institutional editorials, & letters-to-the-editor), **W\_newsp\_brdsh\_t\_nat\_misc** (broadsheet national newspapers: miscellaneous material), **W\_newsp\_brdsh\_t\_nat\_report** (broadsheet national newspapers: home & foreign news reportage), **W\_newsp\_brdsh\_t\_nat\_science** (broadsheet national newspapers: science material), **W\_newsp\_brdsh\_t\_nat\_social** (broadsheet national newspapers: material on lifestyle, leisure, belief & thought), **W\_newsp\_brdsh\_t\_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	<b>W</b> (written), <b>S</b> (spoken)
Author age	<b>0-14 yrs</b> (band 1), <b>15-24 yrs</b> (band 2), <b>25-34 yrs</b> (band 3), <b>35-44 yrs</b> (band 4), <b>45-59 yrs</b> (band 5), <b>60+ yrs</b> (band 6), --- (unclassified)
Author sex	<b>Male, Female, Mixed, Unknown</b> , --- (not applicable/available)
Author type	<b>Corporate, Multiple, Sole, Unknown</b> /unclassified
Audience age	<b>child, teen, adult</b> , --- (unclassified)
Audience sex	<b>male, female, mixed</b> , --- (unclassified)
Audience level	<b>low</b> (level 1), <b>medium</b> (level 2), <b>high</b> (level 3), --- (unclassified)
Sampling	whole text ( <b>whl</b> ), beginning sample ( <b>beg</b> ), middle sample ( <b>mid</b> ), end sample ( <b>end</b> ), composite ( <b>cmp</b> ), unknown/not applicable (--).
Circulation Status	(formerly 'reception status'): <b>Low, Medium, High</b> (blank for unclassified texts)

**Table 1** Information fields and possible values in the [BNCW 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 [BNCW 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'<sup>12</sup>. Or if you are

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

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 [BNCW 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).

<b>BNC World SPOKEN</b>	<b>No. of words</b>	<b>%</b>	<b>Big Genre</b>	<b># of Files</b>
<b>S_brodcast_discussn</b>	757,317	7.3%	Broadcast 10.2%	<b>53</b>
S_brodcast_documentary	41,540	0.4%		10
S_brodcast_news	261,278	2.5%		12
<b>S_classroom</b>	429,970	4.2%		<b>58</b>
S_consult	138,011	1.3%		128
S_conv	4,206,058	40.7%		153
S_courtroom	127,474	1.2%		13
S_demonstratn	31,772	0.3%		6
S_interview	123,816	1.2%		Interviews 13
S_interview_oral_history	815,540	7.9%		9.1% 119
S_lect_commerce	15,105	0.1%	Lectures 2.9%	3
S_lect_humanities_arts	50,827	0.5%		4
S_lect_nat_science	22,681	0.2%		4
S_lect_polit_law_edu	50,881	0.5%		7
S_lect_soc_science	159,880	1.5%		13
S_meeting	1,377,520	13.3%		132
S_parliament	96,239	0.9%		6
S_pub_debate	283,507	2.7%		16
S_sermon	82,287	0.8%		16
S_speech_scripted	200,234	1.9%	Speeches 6.4%	26
S_speech_unscripted	464,937	4.5%	51	
S_sportslive	33,320	0.3%		4
S_tutorial	143,199	1.4%		18
S_unclassified	421,554	4.1%		44
<b>TOTAL</b>	<b>10,334,947</b>	<b>100.00%</b>		<b>909</b>

**Table 2 Breakdown of Spoken BNC World 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. In BNC World, all

except HDE were removed (HDE is not included in the above table). Two other files (from the two genres with numbers in red above) were removed from BNC World. Hence there are 909 genre-categorised spoken files in BNC World instead of the original 915 in the first distribution of the BNC (BNC1).

<b>BNC World WRITTEN</b>	<b>No. of words</b>	<b>%</b>	<b>Big Genre</b>	<b># of Files</b>	
W_ac_humanities_arts	3,321,867	3.8%	Academic Prose 17.7%	87	
W_ac_medicine	1,421,933	1.6%		24	
W_ac_nat_science	1,111,840	1.3%		43	
W_ac_polit_law_edu	4,640,346	5.3%		186	
W_ac_soc_science	4,247,592	4.9%		138	
W_ac_tech_engin	686,004	0.8%	Non-printed essays	23	
W_admin	219,946	0.3%		12	
W_advert	558,133	0.6%		60	
W_biography	3,528,564	4.0%		100	
W_commerce	3,759,366	4.3%		112	
W_email	213,045	0.2%		7	
W_essay_sch	146,530	0.2%		7	
W_essay_univ	65,388	0.1%		4	
W_fict_drama	45,757	0.1%		Fiction 18.6%	2
W_fict_poetry	222,451	0.3%			30
W_fict_prose	15,926,677	18.2%	432		
W_hansard	1,156,171	1.3%	Letters 0.2%	4	
W_institut_doc	546,261	0.6%		43	
W_instructional	436,892	0.5%		15	
W_letters_personal	52,480	0.1%		6	
W_letters_prof	66,031	0.1%		11	
W_misc	9,140,957	10.5%	Broadsheet National Newspapers 3.5%	500	
W_news_script	1,292,156	1.5%		32	
W_newsp_brdsh_t_nat_arts	351,811	0.4%		51	
W_newsp_brdsh_t_nat_commerce	424,895	0.5%		44	
W_newsp_brdsh_t_nat_editorial	101,742	0.1%		12	
W_newsp_brdsh_t_nat_misc	1,032,943	1.2%		95	
W_newsp_brdsh_t_nat_reportage	663,355	0.8%		49	
W_newsp_brdsh_t_nat_science	65,293	0.1%		29	
W_newsp_brdsh_t_nat_social	81,895	0.1%		36	
W_newsp_brdsh_t_nat_sports	297,737	0.3%		24	
W_newsp_other_arts	239,258	0.3%	Regional & Local Newspapers 6.4%	15	
W_newsp_other_commerce	415,396	0.5%		17	
W_newsp_other_report	2,717,444	3.1%		39	
W_newsp_other_science	54,829	0.1%		23	
W_newsp_other_social	1,143,024	1.3%		37	
W_newsp_other_sports	1,027,843	1.2%	9		
W_newsp_tabloid	728,413	0.8%	Tabloids	6	
W_non_ac_humanities_arts	3,751,865	4.3%	Non-academic Prose 19.1%	111	
W_non_ac_medicine	498,679	0.6%		17	
W_non_ac_nat_science	2,508,256	2.9%		62	
W_non_ac_polit_law_edu	4,477,831	5.1%		93	
W_non_ac_soc_science	4,187,649	4.8%		128	
W_non_ac_tech_engin	1,209,796	1.4%		123	

W_pop_lore	7,376,391	8.5%		211
W_religion	1,121,632	1.3%		35
TOTAL	87,284,364	100.0%		3144

**Table 3 Breakdown of Written BNC World Genres**

*N.B.* BNC World has 65 files less than BNC version 1, most of them (53) from the prose fiction genre. Affected genres/figures are in red above.

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.<sup>13</sup> For example, here is how the various subject disciplines are categorised in the LOB corpus and in the [BNCW Index](#):

<b>LOB (&amp; <a href="#">BNCW Index</a>) Category</b>	<b>Subjects/Disciplines Represented</b>
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 [BNCW 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 ‘technology & engineering’ depending on whether they were pure or applied.<sup>14</sup> 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.

<sup>13</sup> 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.

<sup>14</sup> People who disagree with these classifications may use the ‘Keywords’ and ‘Title’ fields to find the relevant files and re-classify them as desired.

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 [BNCW 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 discursively 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 [BNCW Index](#) are advised to read the notes/documentation given here, and to be clear what the various domain and genre labels mean.<sup>15</sup> To 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:

---

<sup>15</sup> The *domain* labels in the [BNCW Index](#) are largely unchanged (i.e. they reflect the decisions of the BNC compilers). Some egregious errors were corrected, however, and reported to the BNC project for fixing in the new release, BNC World Edition.

<i>Text ID</i>	<i>Medium</i>	<i>Domain</i>	<i>Bibliographical Details</i>
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 [BNCW 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:

<i>Text ID</i>	<i>Medium</i>	<i>Domain</i>	<i>Details</i>
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 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 [BNCW Index](#)

The [BNCW Index](#) is distributed in the Microsoft Excel® spreadsheet format as well as in a tab-delimited format (it has also been incorporated into [BNCweb](#) and two custom-built, user-friendly programs: see below).<sup>16</sup> 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\_brdrast\_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.<sup>17</sup> 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<sup>18</sup>. 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 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

---

<sup>16</sup> The [BNCW Index](#) spreadsheet should be referenced as being available from: <http://clix.to/davidlee00>.

<sup>17</sup> 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.**

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

[WordSmith](#) or [MonoConc](#)<sup>19</sup> 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.

## Appendix: 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**. In BNC World, only duplicate file **HDE** remains.
- 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.
- 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.

---

<sup>19</sup>

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 also has more sophisticated sub-corpus querying facilities, including a facility for restricting queries by the genre categories discussed here.

[The following are errors reported by [Sebastian Hoffman](#) to the [BNC-discuss list](#) on 6<sup>th</sup> 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).

**[13 Aug 2002: More BNC World Edition errors spotted by me]**

- 16) **KB7**: social class of respondent (and therefore the file as a whole) should be C2 (not DE).
- 17) Other errors in **domain** (or, for spoken texts, **context** or **social class**) classification are as shown in the following table:

FileID	Official BNC-WE Domain	BNCW Index (=Correct) Domain
A1U	wridom4, W_soc_science	wridom5, W_world_affairs
A29	wridom4, W_soc_science	wridom5, W_world_affairs
A2L	wridom4, W_soc_science	wridom5, W_world_affairs
A2W	wridom8, W_belief_thought	wridom5, W_world_affairs
A39	wridom7, W_arts	wridom9, W_leisure
A3F	wridom4, W_soc_science	wridom5, W_world_affairs
A44	wridom8, W_belief_thought	wridom5, W_world_affairs
A47	wridom4, W_soc_science	wridom5, W_world_affairs
A4G	wridom8, W_belief_thought	wridom5, W_world_affairs
A5B	wridom9, W_leisure	wridom7, W_arts
AC6	wridom1, W_imaginative	wridom4, W_soc_science
AMG	wridom3, W_app_science	wridom4, W_soc_science
B0J	wridom7, W_arts	wridom9, W_leisure
B13	wridom5, W_world_affairs	wridom8, W_belief_thought
CN9	wridom2, W_nat_science	wridom6, W_commerce
D94	scgdom3, S_cg_public_instit	scgdom2, S_cg_business

D95	scgdom3, S_cg_public_instit	scgdom2, S_cg_business
D96	scgdom3, S_cg_public_instit	scgdom2, S_cg_business
D97	scgdom3, S_cg_public_instit	scgdom2, S_cg_business
DCH	scgdom3, S_cg_public_instit	scgdom2, S_cg_business
F7J	scgdom3, S_cg_public_instit	scgdom2, S_cg_business
FPW	wridom8, W_belief_thought	wridom7, W_arts
GX2	wridom3, W_app_science	wridom6, W_commerce
GX8	wridom2, W_nat_science	wridom6, W_commerce
HAU	wridom5, W_world_affairs	wridom6, W_commerce
HDB	wridom1, W_imaginative	wridom9, W_leisure
HDE	scgdom2, S_cg_business	REDUNDANT FILE: all of HDE is included in HEP
HEP	scgdom4, S_cg_leisure	scgdom2, S_cg_business
HEW	scgdom1, S_cg_education	scgdom4, S_cg_leisure
HEY	scgdom1, S_cg_education	scgdom4, S_cg_leisure
HJ9	wridom4, W_soc_science	wridom6, W_commerce
HJB	wridom4, W_soc_science	wridom6, W_commerce
HJL	wridom1, W_imaginative	wridom9, W_leisure
HP5	wridom3, W_app_science	wridom6, W_commerce
HT5	wridom1, W_imaginative	wridom9, W_leisure
HX5	wridom4, W_soc_science	wridom6, W_commerce
J1M	wridom9, W_leisure	wridom5, W_world_affairs
J1S	wridom4, W_soc_science	wridom6, W_commerce
J2A	wridom4, W_soc_science	wridom6, W_commerce
JSE	scgdom1, S_cg_education	scgdom2, S_cg_business
JT5	scgdom1, S_cg_education	scgdom2, S_cg_business
JTE	scgdom3, S_cg_public_instit	scgdom2, S_cg_business
K1B	wridom9, W_leisure	wridom5, W_world_affairs
K1C	wridom9, W_leisure	wridom5, W_world_affairs
K1D	wridom9, W_leisure	wridom5, W_world_affairs
K1E	wridom9, W_leisure	wridom5, W_world_affairs
K1F	wridom9, W_leisure	wridom5, W_world_affairs
K1G	wridom9, W_leisure	wridom5, W_world_affairs
K1H	wridom9, W_leisure	wridom5, W_world_affairs
K1J	wridom9, W_leisure	wridom5, W_world_affairs
K1K	wridom9, W_leisure	wridom5, W_world_affairs
K1L	wridom9, W_leisure	wridom5, W_world_affairs
K1M	wridom9, W_leisure	wridom5, W_world_affairs
K1N	wridom9, W_leisure	wridom5, W_world_affairs
K1P	wridom9, W_leisure	wridom5, W_world_affairs
K1R	wridom9, W_leisure	wridom5, W_world_affairs
K1S	wridom9, W_leisure	wridom5, W_world_affairs
K1T	wridom9, W_leisure	wridom5, W_world_affairs
K1U	wridom9, W_leisure	wridom5, W_world_affairs
K1V	wridom9, W_leisure	wridom5, W_world_affairs
K1W	wridom9, W_leisure	wridom5, W_world_affairs
K1X	wridom9, W_leisure	wridom5, W_world_affairs
K1Y	wridom9, W_leisure	wridom5, W_world_affairs
K20	wridom9, W_leisure	wridom5, W_world_affairs
K21	wridom9, W_leisure	wridom5, W_world_affairs
K22	wridom9, W_leisure	wridom5, W_world_affairs
K23	wridom9, W_leisure	wridom5, W_world_affairs
K24	wridom9, W_leisure	wridom5, W_world_affairs

K25	wridom9, W_leisure	wridom5, W_world_affairs
K26	wridom9, W_leisure	wridom5, W_world_affairs
K27	wridom9, W_leisure	wridom5, W_world_affairs
K28	wridom9, W_leisure	wridom5, W_world_affairs
K6C	scgdom4, S_cg_leisure	scgdom1, S_cg_education
K6D	scgdom4, S_cg_leisure	scgdom1, S_cg_education
K6E	scgdom4, S_cg_leisure	scgdom1, S_cg_education
K6F	scgdom4, S_cg_leisure	scgdom1, S_cg_education
K6G	scgdom4, S_cg_leisure	scgdom1, S_cg_education
KA1	wridom9, W_leisure	wridom7, W_arts
KA4	wridom5, W_world_affairs	wridom9, W_leisure
KB7	sdecla4, S_Demog_DE	sdecla3, S_Demog_C2
KRW	scgdom1, S_cg_education	scgdom4, S_cg_leisure
KRX	scgdom1, S_cg_education	scgdom3, S_cg_public_instit
KS2	scgdom1, S_cg_education	scgdom4, S_cg_leisure

### [30 Oct 2002: More BNC World Edition errors spotted by me]

Changed 'Interaction Type' from 'monologue' to 'dialogue' for the following files: **HEW** and **KJT** ('live' sports commentaries). They predominantly consist of casual chat and banter between the commentators.

**HPS** probably does not belong to the genre 'W\_essay\_univ' as it is unlikely to have been written by a university student. I have now placed it under 'W\_misc' because it is unpublished. Alternatively, it may be considered under 'W\_non\_ac\_humanities'.

~~ \* ~~

*Last edited: 30 Oct 2002*

*Previous editions: 10 May 2002, 4 May 2002, 25 Oct 2001, 13 Aug 2002*

*Send Feedback, Comments, Error Reports to:* [Please type: **david\_lee00** then add **@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.