2001

Genres, registers, text types, domains and styles: Clarifying the concepts and nevigating a path through the BNC jungle

David Y. W. Lee
University of Wollongong, dlee@uow.edu.au

Publication Details

Lee, D. Y. W.. 2001, 'Genres, registers, text types, domains and styles: Clarifying the concepts and nevigating a path through the BNC jungle', Language Learning and Technology, vol. 5, no. 3, pp. 37-72.

Language Learning & Technology September 2001, Vol. 5, Num. 3.
http://llt.msu.edu/vol5num3/lee/ pp. 37-72

GENRES, REGISTERS, TEXT TYPES, DOMAINS, AND STYLES: CLARIFYING THE CONCEPTS AND NAVIGATING A PATH THROUGH THE BNC JUNGLE

David YW Lee
Lancaster University, UK

ABSTRACT

In this paper, an attempt is first made to clarify and tease apart the somewhat confusing terms *genre, register, text type, domain, sublanguage,* and *style.* The use of these terms by various linguists and literary theorists working under different traditions or orientations will be examined and a possible way of synthesising their insights will be proposed and illustrated with reference to the disparate categories used to classify texts in various existing computer corpora. With this terminological problem resolved, a personal project which involved giving each of the 4,124 British National Corpus (BNC, version 1) files a descriptive "genre" label will then be described. The result of this work, a spreadsheet/database (the "BNC Index") containing genre labels and other types of information about the BNC texts will then be described and its usefulness shown. It is envisaged that this resource will allow linguists, language teachers, and other users to easily navigate through or scan the huge BNC jungle more easily, to quickly ascertain what is there (and how much) and to make informed selections from the mass of texts available. It should also greatly facilitate genre-based research (e.g., EAP, ESP, discourse analysis, lexicogrammatical, and collocational studies) and focus everyday classroom concordancing activities by making it easy for people to restrict their searches to highly specified sub-sets of the BNC using PC-based concordancers such as *WordSmith, MonoConc,* or the Web-based *BNCWeb.*

INTRODUCTION

Most corpus-based studies rely implicitly or explicitly on the notion of genre or the related concepts register, text type, domain, style, sublanguage, message form, and so forth. There is much confusion surrounding these terms and their usage, as anyone who has done any amount of language research knows. The aims of this paper are therefore two-fold. I will first attempt to distinguish among the terms because I feel it is important to point out the different nuances of meaning and theoretical orientations lying behind their use. I then describe an attempt at classifying the 4,124 texts in the British National Corpus (BNC) in terms of a broad sense of genre, in order to give researchers and language teachers a better avenue of approach to the BNC for doing all kinds of linguistic and pedagogical research.

Categorising Texts: Genres, Registers, Domains, Styles, Text Types, & Other Confusions

Why is it important to know what these different terms mean, and why should corpus texts be classified into *genres?* The short answer is that language teachers and researchers need to know exactly what kind of language they are examining or describing. Furthermore, most of the time we want to deal with a specific *genre* or a manageable set of genres, so that we can define the scope of any generalisations we make. My feeling is that *genre* is the level of text categorisation which is theoretically and pedagogically most useful and most practical to work with, although classification by *domain* is important as well (see discussion below). There is thus a real need for large-scale general corpora such as the BNC to clearly label and classify texts in a way that facilitates language description and research, beyond the

very broad classifications currently in place. It is impossible to make many useful generalisations about "the English language" or "general English" since these are abstract constructions. Instead, it is far easier and theoretically more sound to talk about the language of different *genres* of text, or the language(s) used in different *domains,* or the different types of *register* available in a language, and so forth. Computational linguists working in areas of natural language processing/language engineering have long realised the need to target the scope of their projects to very specific areas, and hence they talk about *sublanguages* such as air traffic control talk, journal articles on lipoprotein kinetics, navy telegraphic messages, weather reports, and aviation maintenance manuals. (see Grishman & Kittredge, 1986; Kittredge & Lehrberger, 1982, for detailed discussions of "sublanguages").

The terminological issue I grapple with here is a very vexing one. Although not all linguists will recognise or actively observe the distinctions I am about to make (in particular, the use of the term *text type*, which can be used in a very vague way to mean almost anything), I believe there is actually more consensus on these issues than users of these terms themselves realise, and I hope to show this below.

Internal Versus External Criteria: Text Type & Genre

One way of making a distinction between *genre* and *text type* is to say that the former is based on external, non-linguistic, "traditional" criteria while the latter is based on the internal, linguistic characteristics of texts themselves (Biber, 1988, pp. 70 & 170; EAGLES, 1996).[1] A *genre*, in this view, is defined as a category assigned on the basis of **external** criteria such as intended audience, purpose, and activity type, that is, it refers to a conventional, culturally recognised grouping of texts based on properties other than lexical or grammatical (co-)occurrence features, which are, instead, the **internal** (linguistic) criteria forming the basis of *text type* categories. Biber (1988) has this to say about external criteria:

> Genre categories are determined on the basis of external criteria relating to the speaker's purpose and topic; they are assigned on the basis of use rather than on the basis of form. (p. 170)

However, the EAGLES (1996)[2] authors would quibble somewhat with the inclusion of the word *topic* above and argue that one should not think of topic as being something to be established a priori, but rather as something determined on the basis of internal criteria (i.e., linguistic characteristics of the text):

> Topic is the lexical aspect of internal analysis of a text. Externally the problem of classification is that there are too many possible methods, and no agreement or stability in societies or across them that can be built upon ... The boundaries between ... topics are ultimately blurred, and we would argue that in the classification of topic for corpora, it is best done on a higher level, with few categories of topic which would alter according to the language data included. There are numerous ways of classifying texts according to topic. Each corpus project has its own policies and criteria for classification … The fact that there are so many different approaches to the classification of text through topic, and that different classificatory topics are identified by different groups indicates that existing classification[s] are not reliable. They do not come from the language, and they do not come from a generally agreed analysis. However they are arrived at, they are subjective, and … the resulting typology is only one view of language, among many with equal claims to be the basis of a typology. (p. 17)

So perhaps it is best to disregard the word "topic" in the quote from Biber above, and take *genres* simply as categories chosen on the basis of fairly easily definable external parameters. Genres also have the property of being recognised as having a certain legitimacy as groupings of texts within a speech community (or by sub-groups within a speech community, in the case of specialised genres). This is

essentially the view of *genre* taken by Swales (1990, pp. 24-27), who talks about genres being "owned" (and, to varying extents, policed) by particular discourse communities.

Without going into the minutiae of the EAGLES' recommendations, all I will say is that detailed, explicit recommendations do not yet exist in terms of identifying *text types* or, indeed, any so-called "internal criteria." That is, there are as yet, no widely-accepted or established text-type-based categories consisting of texts which cut across traditionally recognisable genres on the basis of internal linguistic features (see discussion below). On the subject of potentially useful internal classificatory criteria, the EAGLES authors mention the work of Phillips (1983) under the heading of *topic* (the "aboutness" or "intercollocation of collocates" or "lexical macrostructures" of texts), and the work of Biber (1988, 1989) and Nakamura (1986, 1987, 1992, 1993) under the heading of *style* (which the EAGLES' authors basically divide into "formal/informal," combining this with parameters such as "considered/impromptu" and "one-way/interactive"). However, the authors offer no firm recommendations, merely the observation that "these are only shafts of light in a vast darkness" (p. 25), and they do not mention what a possible *text type* could be (in fact, no examples are even given of possible labels for text types). At present, all corpora use only external criteria to classify texts. Indeed, as Atkins, Clear, & Ostler (1992, p. 5) note, there is a good reason for this:

> The initial selection of texts for inclusion in a corpus will inevitably be based on external evidence primarily … A corpus selected entirely on internal criteria would yield no information about the relation between language and its context of situation.

The EAGLES (1996) authors add that

> [the] classification of texts based purely on internal criteria does not give prominence to the sociological environment of the text, thus obscuring the relationship between the linguistic and non-linguistic criteria. (p. 7)

Coming back to the distinction between *genre* and *text type,* therefore, the main thing to remember here is what the two different approaches to classification mean for texts and their categorisation. In theory, two texts may belong to the same *text type* (in Biber's sense) even though they may come from two different *genres* because they have some similarities in linguistic form (e.g., biographies and novels are similar in terms of some typically "past-tense, third-person narrative" linguistic features). This highly restricted use of *text type* is an attempt to account for variation **within** and **across** genres (and hence, in a way, to go "above and beyond" genre in linguistic investigations). Biber's (1989, p. 6) use of the term, for example, is prompted by his belief that "genre distinctions do not adequately represent the underlying text types of English …; linguistically distinct texts within a genre represent different text types; linguistically similar texts from different genres represent a single text type."

Paltridge (1996), in an article on "Genre, Text Type, and the Language Learning Classroom," makes reference to Biber (1988; but, crucially, not to Biber 1989)[3] and proposes a usage of the terms *genre* and *text type* which he claims is in line with Biber's external/internal distinction, as delineated above. It is clear from the article, however, that what Paltridge means by "internal criteria" differs considerably from what Biber meant. Paltridge proposes the following distinction:

Table 1. Paltridge's Examples of Genres and "Text Types" (based on Hammond, Burns, Joyce, Brosnan, & Gerot, 1992)

Genre	Text Type
Recipe	Procedure
Personal letter	Anecdote
Advertisement	Description
Police report	Description
Student essay	Exposition
Formal letter	Exposition
Format letter	Problem–Solution
News item	Recount
Health brochure	Procedure
Student assignment	Recount
Biology textbook	Report
Film review	Review

As can be seen, what Paltridge calls "text types" are probably better termed "discourse/rhetorical structure types," since the determinants of his "text types" are not surface-level lexicogrammatical or syntactic features (Biber's "internal linguistic features"), but rhetorical patterns (which is what Hoey, 1986, p. 130, for example, calls them). Paltridge's sources, Meyer (1975), Hoey (1983), Crombie (1985) and Hammond et al. (1992) are all similarly concerned with text-level/discoursal/rhetorical structures or patterns in texts, which most linguists would probably not consider as constituting 'text types' in the more usual sense.

Returning to Biber's distinction between *genre* and *text type*, then, what we can say is that his "internal versus external" distinction is attractive. However, as noted earlier, the main problem is that linguists have still not firmly decided on or enumerated or described in concrete terms the kinds of text types (in Biber's sense) we would profit from looking at. Biber's (1989) work on text typology (see also Biber & Finegan,1986) using his factor-analysis-based multi-dimensional (MD) approach is the most suggestive work so far in this area, but his categories do not seem to have been taken up by other linguists. His eight text types (e.g., "informational interaction," "learned exposition," "involved persuasion") are claimed to be maximally distinct in terms of their linguistic characteristics. The classification here is at the level of individual texts, not groups such as "genres," so texts which nominally "belong together" in a "genre" (in terms of external criteria) may land up in different text types because of differing linguistic characteristics. An important caveat to mention, however, is that there are many questions surrounding the statistical validity, empirical stability, and linguistic usefulness of the linguistic "dimensions" from which Biber derives these "text types," or clusters of texts sharing internal linguistic characteristics (see Lee, 2000, for a critique) and hence these text typological categories should be taken as indicative rather than final. Kennedy (1998) has said, for example, that

> Some of the text types established by the factor analysis do not seem to be clearly different from each other. For example, the types "learned" and "scientific" exposition … may differ only in some cases because of a higher incidence of active verbs in the "learned" text type. (p. 188)

One could also question the aptness or helpfulness of some of the text type labels (e.g., how useful is it to know that 29% of "official documents" belong to the text type "scientific exposition"?).

It therefore still remains to be seen if stable and valid dimensions of (internal) variation, which can serve as useful criteria for text typology, can be found. At the risk of rocking the boat, I would also like to say that, personally, I am not convinced that there is a pressing need to determine "all the text types in the

English language" or to balance corpora on the basis of these types. Biber (1993) notes that it is more important as a first step in compiling a corpus to focus on covering all the situational parameters of language variation, because they can be determined prior to the collection of texts, whereas

> there is no a priori way to identify linguistically defined types ... [however,] the results of previous research studies, as well as on-going research during the construction of a corpus, can be used to assure that the selection of texts is *linguistically* as well as *situationally representative* [italics added]. (p. 245)

My question, however, is: what does it mean to say that a corpus is "linguistically representative" or linguistically balanced? Also, why should this be something we should strive towards? The EAGLES' (1996) authors say that we should see progress in corpus compilation and text typology as a cyclical process:

> The internal linguistic criteria of the text [are] analysed subsequent to the initial selection based on external criteria. The linguistic criteria are subsequently upheld as particular to the genre … [Thus] classification begins with external classification and subsequently focuses on linguistic criteria. If the linguistic criteria are then related back to the external classification and the categories adjusted accordingly, a sort of cyclical process ensues until a level of stability is established. (p. 7)

Or, as the authors say later, this process is one of "frequent cross-checking between internal and external criteria so that each establishes a framework of relevance for the other" (p. 25). Beyond these rather abstract musings, however, there is not enough substantive discussion of what *text types* or other kinds of internally-based criteria could possibly look like or how exactly they would be useful in balancing corpora.

In summary, with *text type* still being an elusive concept which cannot yet be established explicitly in terms of linguistic features, perhaps the looser use of the term by people such as Faigley and Meyer (1983) may be just as useful: they use *text type* in the sense of the traditional four-part rhetorical categories of *narrative, description, exposition* and *argumentation.* Steen (1999, p. 113) similarly calls these four classes "types of discourse."[4] Stubbs (1996, p. 11), on the other hand, uses *text type* and *genre* interchangeably, in common, perhaps, with most other linguists. At present, such usages of *text type* (which do not observe the distinctions Biber and EAGLES try to make) are perhaps as consistent and sensible as any, as long as people make it clear how they are using the terms. It does seem redundant, however, to have two terms, each carrying its own historical baggage, both covering the same ground.

"Genre," "Register," and "Style"

Other terms often used in the literature on language variation are *register* and *style.* I will now walk into a well-known quagmire and try to distinguish between the terms *genre, register,* and *style.* In his *Dictionary of Linguistics and Phonetics,* Crystal (1991, p. 295) defines *register* as "a variety of language defined according to its use in social situations, e.g. a register of scientific, religious, formal English." (Presumably these are three different registers.) Interestingly, Crystal does not include *genre* in his dictionary, and therefore does not try to define it or distinguish it from other similar/competing terms. In Crystal & Davy (1969), however, the word *style* is used in the way most other people use *register:* to refer to particular ways of using language in particular contexts. The authors felt that the term *register* had become too loosely applied to almost any situational variety of language of any level of generality or abstraction, and distinguished by too many different situational parameters of variation. (Using *style* in the same loose fashion, however, hardly solves anything, and, as I argue below, goes against the usage of *style* by most people in relation to individual texts or individual authors/speakers.)

The two terms *genre*[5] and *register* are the most confusing, and are often used interchangeably, mainly because they overlap to some degree. One difference between the two is that *genre* tends to be associated

more with the organisation of culture and social purposes around language (Bhatia, 1993; Swales, 1990), and is tied more closely to considerations of ideology and power, whereas *register* is associated with the organisation of situation or immediate context. Some of the most elaborated ideas about *genre* and *register* can be found within the tradition of systemic functional grammar. The following diagram (Martin & Matthiessen, 1991, reproduced in Martin, 1993, p. 132), shows the relation between language and context, as viewed by most practitioners of systemic-functional grammar:

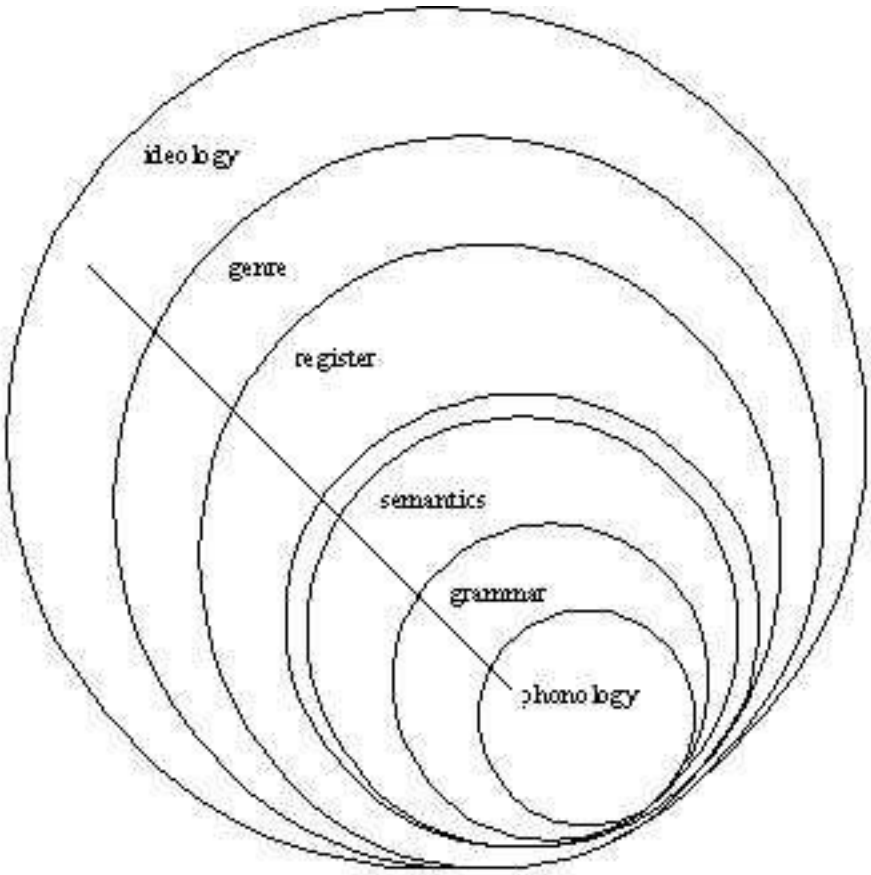

Figure 1. Language and context in the systemic functional perspective

In this tradition, *register* is defined as a particular configuration of *field*, *tenor*, and *mode* choices (in Hallidayan grammatical terms), in other words, a language variety functionally associated with particular contextual or situational parameters of variation and defined by its linguistic characteristics. The following diagram illustrates this more clearly:

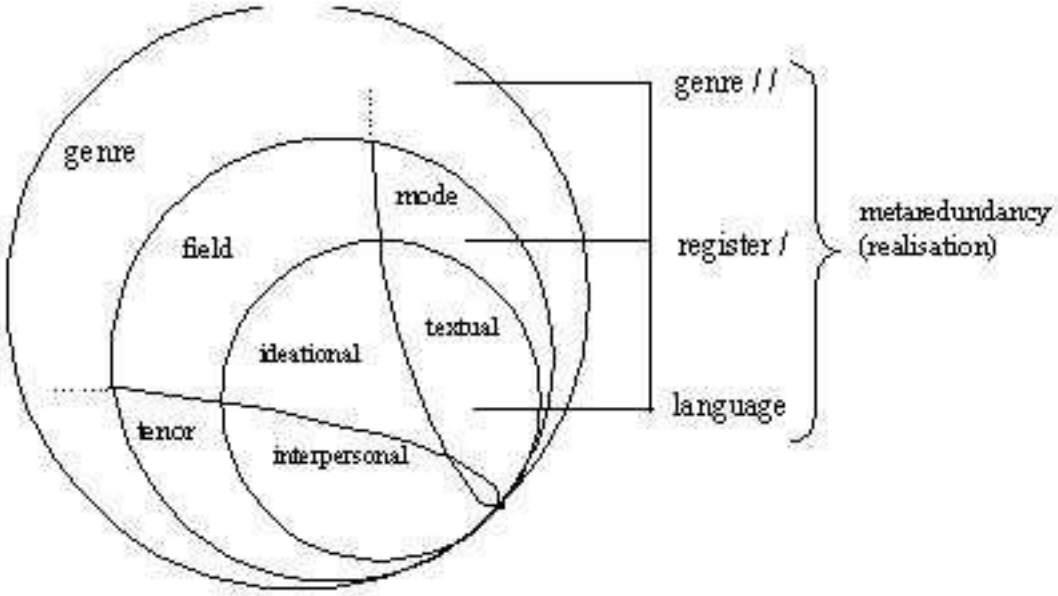

Figure 2. Metafunctions in relation to register and genre[6]

Genre, on the other hand, is more abstractly defined:

> A genre is known by the meanings associated with it. In fact the term "genre" is a short form for the more elaborate phrase "genre-specific semantic potential" … Genres can vary in delicacy in the same way as contexts can. But for some given texts to belong to one specific genre, their structure should be some possible realisation of a given GSP Generic Structure Potential … It follows that texts belonging to the same genre can vary in their structure; the one respect in which they cannot vary without consequence to their genre-allocation is the obligatory elements and dispositions of the GSP. (Halliday & Hasan, 1985, p. 108)

> [T]wo layers of context are needed -- with a new level of *genre* [italics added] posited above and beyond the field, mode and tenor register variables … Analysis at this level has concentrated on making explicit just which combinations of field, tenor and mode variables a culture enables, and how these are mapped out as *staged, goal-oriented social processes* [italics added]. (Eggins & Martin, 1997, p. 243)

These are rather theory-specific conceptualisations of *genre*, and are therefore a little opaque to those not familiar with systemic-functional grammar. The definition of *genre* in terms of "staged, goal-oriented social **processes**" (in the quote above, and in Martin, Christie, & Rothery, 1987), is, in particular, slightly confusing to those who are more concerned (or familiar) with genres as products (i.e., groupings of texts).

Ferguson (1994), on the other hand, offers a less theory-specific discussion. However, he is rather vague, and talks about (and around) the differences between the two terms while never actually defining them precisely: He seems to regard *register* as a "communicative situation that recurs regularly in a society" (p. 20) and *genre* as a "message type that recurs regularly in a community" (p. 21). Faced with such comparable definitions, readers will be forgiven for becoming a little confused. Also, is *register* only a "communicative situation," or is it a variety of language as well? In any case, Ferguson also seems to equate *sublanguage* with *register* (p. 20) and offers many examples of *registers* (e.g., cookbook recipes, stock market reports, regional weather forecasts) and *genres* (e.g., chat, debate, conversation, recipe, obituary, scientific textbook writing) without actually saying why any of the registers cannot also be

thought of as genres or vice versa. Indeed, sharp-eyed readers will have noted that recipes are included under both *register* and *genre*.

Coming back to the systemic-functional approach, it will be noted that even among subscribers to the "genre-based" approach in language pedagogy (Cope & Kalantzis, 1993), opinions differ on the definition and meaning of *genre.* For J. R. Martin, as we have seen, genre is above and beyond register, whereas for Gunther Kress, *genre* is only one part of what constitutes his notion of *register* (a superordinate term). The following diagram illustrates his use of the terms:

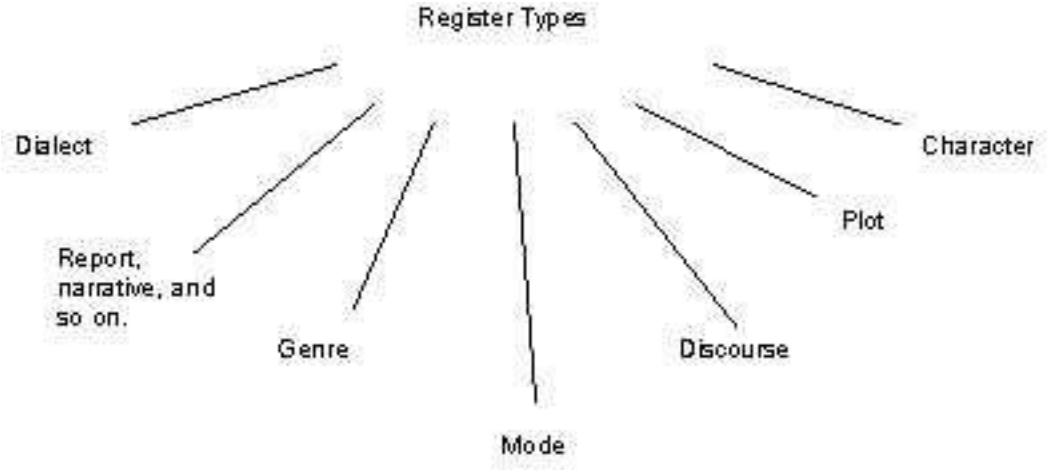

Figure 3. Elements of the composition of text (Kress, 1993, p. 35)

Kress (1993) appears to dislike the fact that *genre* is made to carry too much baggage or different strands of information:

> There is a problem in using such a term [genre] with a meaning that is relatively uncontrollable. In literary theory, the term has been used with relative stability to describe formal features of a text -- epitaph, novel, sonnet, epic -- although at times content has been used to provide a name, [e.g.] epithalamion, nocturnal, alba. In screen studies, as in cultural studies, labels have described both form and content, and at times other factors, such as aspects of production. Usually the more prominent aspect of the text has provided the name. Hence "film noir"; "western" or "spaghetti western" or "psychological" or "Vietnam western"; "sci-fi"; "romance"; or "Hollywood musical"; and similarly with more popular print media. (pp. 31-2)

In other words, Kress is complaining about the fact that

> a great complex of factors is condensed and compacted into the term -- factors to do with the relations of producer and audience, modes of production and consumption, aesthetics, histories of form and so on. (p. 32)

He claims that many linguists, educators, and literacy researchers, especially those working within the Australian-based "genre theory/school" approach, use the term in the same all-encompassing way. Also, he is concerned that the work of influential people like Martin and Rothery has been focussed too much on presenting ideal generic texts and on the successive "unfolding" of "sequential stages" in texts (which are said to reflect the social tasks which the text producers perform; Paltridge, 1995, 1996, 1997):

> The process of classification … seems at times to be heading in the direction of a new formalism, where the 'correct' way to write [any particular text] is presented to students in the form of generic models and exegeses of schematic structure. (Kress, 1993, p. 12)

Those familiar with Kress' work in critical discourse analysis (e.g., Kress & Hodge, 1979) should not be surprised to learn, however, that in his approach to *genre* the focus is instead:

> … on the structural features of the *specific social occasion* in which the text has been produced [, seeing] these as giving rise to particular configurations of linguistic factors in the text which are realisations of, or reflect, these *social relations and structures* [,…e.g.] who has the *power* to initiate turns and to complete them, and how *relations of power* are realised linguistically. In this approach "genre" is a term for only a part of textual structuring, namely the part which has to do with the structuring effect on text of sets of *complex social relations* between consumers and producers of texts. [all italics added] (p. 33)

As can be seen, therefore, there is a superficial terminological difference in the way *genre* is used by some theorists, but no real, substantive disagreement because they both situate it within the broader context of situational and social structure. While *genre* encompasses *register* and goes "above and beyond" it in Martin's (1993, Eggins & Martin, 1997) terms, it is only one component of the larger overarching term *register* in Kress' approach. My own preferred usage of the terms comes closest to Martin's, and will be described below. Before that, however, I will briefly consider two other attempts at clearing up the terminological confusion.

Sampson (1997) calls for re-definitions of *genre, register,* and *style* and the relationships among them, but his argument is not quite lucid or convincing enough. In particular, his proposal for *register* to be recognised as fundamentally to do with an individual's idiolectal variation seems to go against the grain of established usage, and is unlikely to catch on. Biber (Finegan & Biber, 1994, pp. 51-53; 1995, pp. 7-10) does a similar survey, looking at the use of the terms *register, genre, style, sublanguage,* and *text type* in the sociolinguistic literature, and despairingly comes to the conclusion that *register* and *genre,* in particular, cannot be teased apart. He settles on *register* as "the general cover term associated with all aspects of variation in use" (1995, p. 9), but in so doing reverses his choice of the term *genre* in his earlier studies, as in Biber (1988) and Biber & Finegan (1989). (Further, as delineated in Finegan & Biber, 1994, Biber also rather controversially sees register variation as a very fundamental basis or cause of social dialect variation.)

While hoping not to muddy the waters any further, I shall now attempt to state my position on this terminological issue. My own view is that *style* is essentially to do with an individual's use of language. So when we say of a text, "It has a very informal style," we are characterising not the *genre* to which it belongs, but rather the text producer's use of language in that particular instance (e.g., "It has a very quirky style"). The EAGLES (1996) authors are not explicit about their stand on this point, but say they use *style* to mean:

> the way texts are internally differentiated other than by topic; mainly by the choice of the presence or absence of some of a large range of structural and lexical features. Some features are mutually exclusive (e.g. verbs in the active or passive mood), and some are preferential, e.g. politeness markers and mitigators. (p. 22)

As noted earlier, the main distinction they recommend for the stylistic description of corpus texts is *formal/informal* in combination with parameters such as the level of preparation (considered/impromptu), "communicative grouping" (conversational group; speaker/writer and audience; remote audiences) and "direction" (one-way/interactive). This chimes with my suggestion that we should use the term *style* to characterise the internal properties of individual texts or the language use by individual authors, with "formality" being perhaps the most important and fundamental one. Joos's (1961) five famous epithets "frozen," "formal," "informal," "colloquial," and "intimate" come in handy here, but these are only suggestive terms, and may be multiplied or sub-divided endlessly, since they are but five arbitrary points on a sliding scale. On a more informal level, we may talk about speakers or writers having a "humorous,"

"ponderous," or "disjointed" style, or having a "repertoire of styles." Thus, describing one text as "informal" in style is not to say the speaker/writer cannot also write in a "serious' style," even within the same *genre.*

The two most problematic terms, *register* and *genre,* I view as essentially two different points of view covering the same ground. In the same way that any stretch of language can simultaneously be looked at from the point of view of *form* (or category), *function,* or *meaning* (by analogy with the three sides of a cube), *register* and *genre* are in essence two different ways of looking at the same object.[7] *Register* is used when we view a text as language: as the instantiation of a conventionalised, functional configuration of language tied to certain broad societal situations, that is, variety according to use. Here, the point of view is somewhat static and uncritical: different situations "require" different configurations of language, each being "appropriate" to its task, being maximally "functionally adapted" to the immediate situational parameters of contextual use. *Genre* is used when we view the text as a member of a category: a culturally recognised artifact, a grouping of texts according to some conventionally recognised criteria, a grouping according to purposive goals, culturally defined. Here, the point of view is more dynamic and, as used by certain authors, incorporates a critical linguistic (ideological) perspective: Genres are categories established by consensus within a culture and hence subject to change as generic conventions are contested/challenged and revised, perceptibly or imperceptibly, over time.

Thus, we talk about the existence of a *legal register* (focus: language), but of the instantiation of this in the *genres* of "courtroom debates," "wills" and "testaments," "affidavits," and so forth (focus: category membership). We talk about a *formal register,* where "official documents" and "academic prose" are possible exemplar *genres.* In contrast, there is no *literary register,* but, rather, there are literary *styles* and literary *genres,* because the very essence of imaginative writing is idiosyncrasy or creativity and originality (focus on the individual style). My approach here thus closely mirrors that of Fairclough (2000, p. 14) and Eggins & Martin (1997). The latter say that "the linguistic features selected in a text will encode contextual dimensions, both of its immediate context of production (i.e., register) and of its generic identity (i.e., genre), what task the text is achieving in the culture" (p. 237), although they do not clearly set out the difference in terms of a difference in point of view, as I have done above. Instead, as we have seen, they attempt in rather vague terms to define *register* as a variety "organised by metafunction" (Field, Tenor, Mode) and *genre* as something "above and beyond metafunctions." In Biber's (1994) survey of this area of terminological confusion, he mentions the use of terminology by Couture (1986), but fails to note a crucial distinction apparently made by the author:

> Couture's examples of genres and registers seem to be more clearly distinguished than in other studies of this type. For example, *registers* include the *language* used by preachers in sermons, the *language* used by sports reporters in giving a play-by-play description of a football game, and the *language* used by scientists reporting experimental research results. *Genres* include both literary and non-literary *text varieties*, for example, short stories, novels, sonnets, informational reports, proposals, and technical manual. [all italics added] (Finegan & Biber, 1994, p. 52)

Biber does not point out that a key division of labour between the two terms is being made here which has nothing to do with the particular examples of activity types, domains, topics, and so forth: whenever *register* is used, Couture is talking about "the language used by…", whereas when *genre* is used, we are dealing **with "text varieties"** (i.e., groupings of texts).

I contend that it is useful to see the two terms *genre* and *register* as really two different angles or points of view, with *register* being used when we are talking about lexico-grammatical and discoursal-semantic patterns associated with situations (i.e., linguistic patterns), and *genre* being used when we are talking about memberships of culturally-recognisable categories. Genres are, of course, instantiations of registers (each genre may invoke more than one register) and so will have the lexico-grammatical and discoursal-

semantic configurations of their constitutive registers, in addition to specific generic socio-cultural expectations built in.

Genres can come and go, or change, being cultural constructs which vary with the times, with fashion, and with ideological movements within society. Thus, some sub-genres of "official documents" in English have been observed to have changed in recent times, becoming more conversational, personal, and familiar, sometimes in a deliberate way, with manipulative purposes in mind (Fairclough 1992). The genres have thus changed in terms of the *registers invoked* (an aspect of *intertextuality),* among other changes, but the *genre labels* stay the same, since they are descriptors of socially constituted, functional categories of text.

Much of the confusion comes from the fact that language itself sometimes fails us, and we end up using the same words to describe both language (*register* or *style)* and category *(genre).* For example, "conversation" can be a register label ("he was talking in the conversational register"), a style label ("this brochure employs a very conversational style"), or a genre label ("the [super-]genre of casual/face-to-face conversations," a *category* of spoken texts). Similarly, weather reports are cited by Ferguson (1994) as forming a register (from the point of view of the *language* being functionally adapted to the situational purpose), but they are surely also a genre (a culturally recognised category of texts). Ferguson gives "obituaries" as an example of a *genre,* but fails to recognise that there is not really a recognisable "*register* of obituaries" only because the actual language of obituaries is not fixed or conventionalised, allowing considerable variation ranging from humorous and light to serious and ponderous.

Couture (1986) also offers an additional angle on the distinction between *register* and *genre:*

> While registers impose explicitness constraints at the level of vocabulary and syntax, genres impose additional explicitness constraints at the discourse level … Both literary critics and rhetoricians traditionally associate genre with a complete, unified textual structure. Unlike register, genre can only be realized in completed texts or texts that can be projected as complete, for a genre does more than specify kinds of codes extant in a group of related texts; it specifies conditions for beginning, continuing, and ending a text. (p.82)

The important point being made here is that genres are about whole texts, whereas registers are about more abstract, internal/linguistic patterns, and, as such, exist independently of any text-level structures.

In summary, I prefer to use the term *genre* to describe groups of texts collected and compiled for corpora or corpus-based studies. Such groups are all more or less conventionally recognisable as text categories, and are associated with typical configurations of power, ideology, and social purposes, which are dynamic/negotiated aspects of situated language use. Using the term *genre* will focus attention on these facts, rather than on the rather static parameters with which *register* tends to be associated. *Register* has typically been used in a very uncritical fashion, to invoke ideas of "appropriateness" and "expected norms," as if situational parameters of language use have an unquestionable, natural association with certain linguistic features and that social evaluations of contextual usage are given rather than conventionalised and contested. Nevertheless, the term has its uses, especially when referring to that body of work in sociolinguistics which is about "registral variation," where the term tells us we are dealing with *language* varying according to socio-situational parameters. In contrast, the possible parallel term "genre/generic variation" does not seem to be used, because while you can talk about "language variation according to social situations of use," it makes no sense to talk about "categories of texts varying according to the categories they belong to." Of course, I am not saying that genres do not have internal variation (or sub-genres). I am saying that "genre variation" makes no sense as a parallel to "register variation" because while you can talk about language (registers) varying across genres, it is tautologous to talk about genres (text categories) varying across genres or situations. In other words, when we study differences among genres, we are actually studying the way the language varies because of social and

situational characteristics and other genre constraints (registral variation), not the way texts vary because of their categorisation.

Genres as Basic-Level Categories in a Prototype Approach

One problem with genre labels is that they can have so many different levels of generality. For example, some genres such as "academic discourse" are actually very broad, and texts within such a high-level genre category will show considerable internal variation: that is, individual texts within such a genre can differ significantly in their use of language (as, for example, Biber, 1988, has shown). A second problem, as Kress noted, is that different "genres" can be based on so many different criteria (domain, topic, participants, setting, etc.).

There is a possible solution to this. Steen (1999) is an interesting attempt at applying prototype theory (Rosch, 1973a, 1973b, 1978; Taylor, 1989) to the conceptualisation of *genre* (and hence to the formalisation of a taxonomy of discourse; cf. also Paltridge, 1995, who made a similar argument but from a different perspective). Basically, the prototype approach can be summarised by Table 2 (which represents my understanding of Steen's ideas; my own suggestions are marked by "?"):

Table 2. A Prototype Approach to *Genre*

SUPERORDINATE	Mammal	Literature ["SUPER-GENRE"?]	Advertising ["SUPER-GENRE"]
BASIC-LEVEL	Dog/Cat	Novel, Poem, Drama [GENRE]	Advertisement [GENRE]
SUBORDINATE [PROTOTYPE]	Cocker spaniel / Siamese	Western, Romance, Adventure [SUB-GENRE]	Print ad, Radio ad, TV ad, T-shirt ad [SUB-GENRE]

Basic-level categories are those which are in the middle of a hierarchy of terms. They are characterised as having the maximal clustering of humanly-relevant properties (attributes), and are thus distinguishable from superordinate and subordinate terms: "It is at the basic level of categorization that people conceptualize things as perceptual and functional gestalts" (Taylor, 1989, p. 48). A basic-level category, therefore, is one for which human beings can easily find prototypes or exemplars, as well as less prototypical members. Subordinate-level categories, therefore, operate in terms of prototypes or fuzzy boundaries: some are better members than others, but all are valid to some degree because they are cognitively salient along a sliding scale. We can also extend this fuzzy-boundary approach to the other levels (basic-level and superordinate) to account for all kinds of mixed genres and super-genres (e.g., to what degree can Shakespeare's *dramas* be said to be different from *poetry?* When does good *advertising* become a form of *literature* or vice versa?).

Steen (1999) applies the idea of basic-level categories and their prototypes to the conceptualisation of *genre* as follows:

> It is presumably the level of *genre* that embodies the basic level concepts, whereas *subgenres* are the conceptual *subordinates*, and more abstract *classes of discourse* are the *superordinates*. Thus the *genre* of an advertisement is to be contrasted with that of a sermon, a recipe, a poem, and so on. These genres differ from each other on a whole range of attributes … The *subordinates* of the genre of the advertisement are less distinct from each other. The press advertisement, the radio commercial, the television commercial, the Internet advertisement, and so on, are mainly distinguished by *one* feature: their medium. The *superordinate* of the genre of the ad, advertising, is also systematically distinct from the other superordinates by means of only *one* principal attribute, the one of domain: It is "business" for advertising, but it exhibits the respective

> values of "religious", "domestic" and "artistic" for the other examples. [all italics added] (p. 112)

Basically, Steen is proposing that we can recognise genres by their cognitive basic-level status: True genres, being basic-level, are maximally distinct from one another (in terms of certain "attributes" to be discussed below), whereas members at the level of sub-genre (which operate on a prototype basis) or "super-genre"[8] have fewer distinctions among themselves.

The proposal is for genres to be treated as basic-level categories which are characterised by (provisionally) a set of seven attributes: *domain* (e.g., art, science, religion, government), *medium* (e.g., spoken, written, electronic), *content* (topics, themes), *form* (e.g., generic superstructures, à la van Dijk (1985), or other text-structural patterns), *function* (e.g., informative, persuasive, instructive), *type* (the rhetorical categories of "narrative," "argumentation," "description," and "exposition") and *language* (linguistic characteristics: register/style[?]). Steen offers only a preliminary sketch of this approach to genre (and hence to a taxonomy of discourse), and, as it stands, it appears to be too biased towards written genres. Other attributes can (and should) be added: for example, *setting or activity type*, to distinguish a broadcast interview from a private interview; or *audience level*, to distinguish public lectures from university lectures (and both attributes to distinguish the latter from school classroom lessons). Another point is that dependencies among the attributes exist (many values for *domain*, *medium*, and *content* are typically co-selected, for instance). Nevertheless, the approach looks like a promising one, and when fully developed will help us sort out genres from sub-genres.

"GENRES" IN CORPORA

Applying this "fuzzy categories" way of looking at genre to corpus studies, we can see that the categories to which texts have been assigned in existing corpora are sometimes genres, sometimes sub-genres, sometimes "super-genres" and sometimes something else altogether. (This is undoubtedly why the catch-all term "text category" is used in the official documentation for the LOB and ICE-GB corpora. Most of these "text categories" are equivalent to what I am calling "genres" in the BNC Index.) For example, consider ICE-GB corpus categories in Table 3.

Table 3. Text Categories in ICE-GB (figures in parentheses indicate the number of 2,000-word texts in each category)

Medium I	Medium II (?) or Interaction Type (?)	Super-genre or Function	Genres or Sub-genres
SPOKEN (300)	Dialogue (180)	Private (100)	face-to-face conversations (90) phone calls (10)
		Public (80)	classroom lessons (20) broadcast discussions (20) broadcast interviews (10) parliamentary debates (10) legal cross-examinations (10) business transactions (10)
	Monologue (100)	Unscripted (70)	spontaneous commentaries (20) unscripted speeches (30) demonstrations (10) legal presentations (10)
		Scripted (30)	broadcast talks (20) non-broadcast speeches (10)
	Mixed (20)		broadcast news (20)

WRITTEN (200)	Non-Printed (50)	Non-professional writing (20)	student essays (10) student examination scripts (10)
		Correspondence (30)	social letters (15) business letters (15)
	Printed (150)	Academic writing (40)	humanities (10) social sciences (10) natural sciences (10) technology (10)
		Non-academic writing (40)	humanities (10) social sciences (10) natural sciences (10) technology (10)
		Reportage (20)	press news reports (20)
		Instructional writing (20)	administrative/regulatory (10) skills/hobbies (10)
		Persuasive writing (10)	press editorials (10)
		Creative writing (20)	novels/stories (20)

The top row of the table is my attempt at describing what attribute(s) or levels the terms within each column represent. The terms within the last column are what end-users of the corpus normally work with, and can be seen to be either genres or sub-genres, viewed from a prototype perspective (e.g., "broadcast interview" is probably best seen as a sub-genre of "interview," differing mainly in terms of the setting, and business letters differ from social letters mainly in terms of *domain).* Most of the terms in the third column can be said to describe "super-genre" or "super-super-genres," with the exception of "instructional writing" and "persuasive writing" (shaded), which seem more like functional labels.[9]

The British National Corpus (BNC), in contrast, has no text categorisation for written texts beyond that of *domain*, and no categorisation for spoken texts except by "context" and demographic/socio-economic classes. The following diagram shows the breakdown of the BNC:

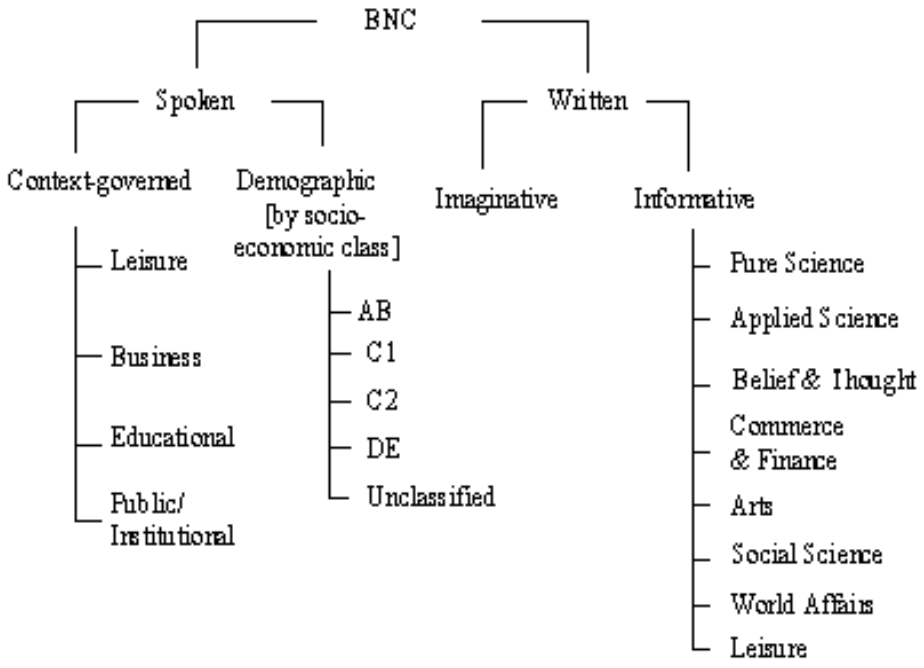

Figure 4. Domains in the British National Corpus (BNC)

It can be seen that for the written texts, *domains* are broad "subject fields" (see Burnard, 1995). These are closely paralleled for the spoken texts by even broader "context" categories covering the major spheres of social life (leisure, business, education, and institutional/public contexts). Apart from considering all the demographically sampled conversations as constituting one super-genre of "casual conversation" and all the written imaginative texts as forming a super-genre "literature," genres cannot easily be found at all under the current domain scheme. More about these BNC categories and their (non-) usefulness will be said in later sections.

Moving on to the LOB corpus (Table 4), we see that it is mostly composed of a mixture of genre and sub-genre labels:

Table 4. Genres in the LOB Corpus

LOB Corpus (Written)
Press: reportage
Press: editorial
Press: reviews
Religion
Skills, trades & hobbies
Popular Lore
Belles lettres, biography, essays
Misc (gov docs, foundation reports, industry reports, college reports, in-house organ)
Learned/scientific writings
General fiction
Mystery & detective fiction
Science fiction
Adventure & western fiction
Romance & love story
Humour

Examined in terms of Steen's genre attributes, the shaded cells in Table 4 above are clearly sub-genres of some general super-genre of "fiction" (both "novels" and "short stories" -- the basic-level genres in Steen's taxonomy -- are included). "Religion," on the other hand, appears to be a *domain* label since it brings together disparate books, periodicals and tracts whose principal common feature is that they are concerned with religion (in this case Christianity).[10]

Why do we have all these different levels or types of categorisation? It is tempting to believe that this is the case because the corpus compilers felt that these were the most useful, salient, or interesting categories -- perhaps these are basic-level genres, or prototypical sub-genres (especially those which keep appearing in different corpora). But is it a problem that the categories differ in terms of their defining attributes and in terms of generality? My personal opinion is that it is not. Cranny-Francis (1993, p. 109) touches on this point and asks:

> If "genre" has this range of different meanings and classificatory procedures -- by formal characteristics, by field -- we might ask what is its value? Why is it so useful to educators, linguists and critics, as well as to publishers, filmmakers, booksellers, readers and viewers?

She suggests that the reason is simply because genre "is never simply formal or semantic [based on field or subject area] and it is not even simply textual." Using the terms as defined in this paper, we could

paraphrase this to read, "genre is never just about situated linguistic patterns *(register),* functional co-occurrences of linguistic features *(text types),* or subject fields *(domain),* and it is not even simply about text-structural/discoursal features (e.g., Martin's [1993] *generic stages,* Halliday & Hasan's [1985] *GSPs,* van Dijk's [1985] *macrostructures,* etc.)." It is, in fact, all of these things. This makes it a messy and complex concept, but it is also what gives it its usefulness and meaningfulness to the average person. They are all genres (whether sub- or super-genres or just plain basic-level genres).

The point of all this is that we need not be unduly worried about whether we are working with genres, sub-genres, domains, and so forth, as long as we roughly know what categories we are working with and find them useful. We have seen that the categories used in various corpora are not necessarily all "proper" genres in a traditional/rhetorical sense or even in terms of Steen's framework, but they can all be seen as "genres" at some level in a fuzzy-category, hierarchical approach. A genre is a basic-level category, which has specified values for most of the seven attributes suggested above and which is maximally distinct from other categories at the same level. "Sub-genres" and "super-genres" are simply other (fuzzy) ways of categorising texts, and have their uses too. The advantages of the prototype approach are that (a) gradience or fuzziness between and within genres is accorded proper theoretical status, and (b) overlapping of categories is not a problem (thus texts can belong to more than one genre).

From one point of view, until we have a clear taxonomy of genres, it may be advisable to put most of our corpus genres in quotation marks, because *genre* is also often used in a folk linguistic way to refer to any more-or-less coherent category of text which a mature, native speaker of a language can easily recognise (e.g., newspaper articles, radio broadcasts), and there are no strict rules as to what level of generality is allowable when recognising genres in this sense. In a prototype approach, however, it does not seriously matter. Some text categories may be based more on the *domain* of discourse (e.g., "business" is a domain label in the BNC for any spoken text produced within a business context, whether it is a committee meeting or a monologic presentation). Spoken texts, which tend to be even more loosely classified in corpus compilations, may simply be categorised on whether they are spontaneous or planned, broadcast or spoken face-to-face, as in the London-Lund Corpus, for instance, which means the categories are "genres" only in a very loose sense. This goes to show that there are still serious issues to grapple with in the conceptualisation of spoken genres (written ones are, in contrast, typically easier to deal with) but that a prototype approach, with its many levels of generality and a set of defining attributes, may help to tighten up our understanding.

These brief visits to the various corpora suggest that there should not be any serious objections (theoretical or otherwise) to the use of the term *genre* to describe most of the corpus categories we have seen. Such usage reflects a looser approach, but there is no requirement for genres to actually be established literary or non-literary genres, only for them to be culturally recognisable as groupings of texts at some level of abstraction. The various corpora also show us that the recognition of genres can be at different levels of generality (e.g., "sermons" vs. "religious discourse"). In the LOB corpus, the category labels appear to be a mix: some are *sub-genre* labels (e.g., "mystery fiction" and "detective fiction"), while others are more properly seen as *domain* labels ("Skills, trades, & hobbies," "Religion"). My own preferred approach with regard to developing a categorisation scheme is to use genre categories where possible, and domain categories where they are more practical (e.g., "Religion"[11]).

THE BNC JUNGLE: THE NEED FOR A PROPER NAVIGATIONAL MAP

Having clarified some of the terminology and concepts and looked at the categories used in a few existing corpora, I want to move on to consider some of the problems with the British National Corpus as it now stands, and then introduce a new resource called the *BNC Index* which (it is hoped) will make it easier for researchers and language learners/teachers to navigate through the numerous texts to find what they need.

Some Existing Problems

Overly Broad Categories. The first problem that prompts the need for a navigational map has to do with the broadness and inexplicitness of the BNC classification scheme. For example, academic and non-academic texts under the domains "Applied Science," "Arts," "Pure/Natural Science," "Social Science," and so forth, are not explicitly differentiated. (It is interesting to note, in this connection, that under the attribute of "genre" in the "text typology" of Atkins et al., 1992, p. 7, no mention is made of the useful distinction between academic and non-academic prose, even though this is employed in one of the earliest corpora, the LOB corpus, where the "learned" category has proved to be among the most popular with linguists.)

Another example that points to the inadequacy of the BNC's categorisation of texts is the way "imaginative" texts are handled. A wide variety of imaginative texts (novels, short stories, poems, and drama scripts) is included in the BNC, which is a good thing because the LOB, for example, does not contain poetry or drama. However, such inclusions are practically wasted if researchers are not actually able to easily retrieve the sub-genres on which they want to work (e.g., poetry) because this information is not recorded in the file headers or in any documentation associated with the BNC. There is at present no way to know whether an "imaginative" text actually comes from a novel, a short story, a drama script or a collection of poems (unless the title actually reflexively includes the words "a novel" or "poems by XYZ"). For example, given text files with titles like "For Now" or "The kiosk on the brink," there is no way of knowing that both of these are actually collections of poems. All the BNC bibliography and file headers tell us is that these are "imaginative" texts, taken from "books."

Classification Errors and Misleading Titles. In the process of some previous research, I found that there were many classificatory mistakes in the BNC (and also in the BNC Sampler): some texts were classified under the wrong category, usually because of a misleading title. For the same reason, even though a limited, computer-searchable bibliographical database of the BNC texts exists[12] (compiled by Adam Kilgarriff), not enough information is included there, and researchers cannot always rely on the titles of the files as indications of their real contents: For example, many texts with "lecture" in their title are actually classroom discussions or tutorial seminars involving a very small group of people, or were popular lectures (addressed to a general audience rather than to students at an institution of higher learning). A good reason for a navigational map, then, is so that we can go beyond the existing information we have about the BNC files (and beyond the mistakes) and to provide genre classifications, so that researchers do not have just the titles of files to go on.

Sub-Genres Within a Single File. Another problem, which will only be touched on briefly because there is no real solution, is that some BNC files are too big and ill-defined in that they contain different genres or sub-genres. For example, newspaper files described in the title as containing "editorial material" include letters-to-the-editor, institutional editorials (those written by the editor), and personal editorials (commentaries/personal columns written by journalists or guest writers), and some courtroom files contain both legal cross-examinations (which are dialogic) as well as legal presentations (summing-up monologues by barristers or judges). This is a problem for lines of linguistic enquiry that rely on relatively homogeneous genres. It is a problem, however, which cannot be solved easily because the splitting of files is beyond the scope of most end-users of the BNC. The problem is just mentioned here as a caution to researchers.

Domains Versus *Genres:* The BNC Sampler & Why We Need Genre Information

The *BNC Users' Reference Guide* states that only three criteria were used to "balance" the corpus: domain, time, and medium. In choosing texts for inclusion into the BNC Sampler (the 2-million word sub-set of the BNC), *domain* was probably the most important criterion used to ensure a wide-enough coverage of a variety of texts. On the BNC Web page for the Sampler, the following comment on its representativeness is made:

> In selecting from the BNC, we tried to preserve the variety of text-types represented, so the Sampler includes in its 184 texts many different *genres* [italics added] of writing and modes of speech.

It should be noted that no real claim to representativeness is made, and that what they really meant was that many different texts were chosen on the basis of *domain* and other criteria.[13] The fact that the Sampler contains many different *genres* is not in doubt, but the texts were not chosen on this basis, since they had no genre classification, and hence the Sampler cannot (and, indeed, it does not) claim to be representative in terms of "genre."

It is my belief that it is because "domain" is such a broad classification in the BNC that the Sampler turned out to be rather unrepresentative of the BNC and of the English language. Anyone wishing to use the Sampler should be under no illusion that it is a balanced corpus or that it represents the full range of texts as in the full BNC. The Sampler may be broadly balanced in terms of the domains, but when broken down by genre, a truer picture emerges of exactly how (un)representative it really is. Appendix A lists missing or unrepresentative genres in the Sampler BNC which demonstrate this.

"Genre" is perhaps a more insightful classification criterion than "domain," as least as far as getting a representatively balanced corpus is concerned. If the compilers of the BNC Sampler had known the genre membership of each BNC text, they would probably have created a more balanced and representative sub-corpus. As things stand, however, any conclusions about "spoken English" or "written English" made on the basis of the BNC Sampler will have to be evaluated very cautiously indeed, bearing in mind the genres missing from the data.

There is another example of how large, undifferentiated categories similar to *domain* can unhelpfully lump disparate kinds of text together. Wikberg (1992) criticises the LOB text category E ("Skills, trades, and hobbies") as being too baggy or eclectic. He demonstrates how, on the evidence of both external and internal criteria, the texts in Category E can actually be better sub-classified into "procedural" versus "non-procedural" discourse. He also notes that it is not just text categories that can be heterogeneous. Sometimes texts themselves are "multitype" or mixed in terms of having different stages with different rhetorical or discourse goals. He thus concludes with the following comment:

> An important point that I have been trying to make is that in the future we need to pay more attention to text theory when compiling corpora. For users of the Brown and the LOB corpora, and possibly other machine-readable texts as well, it is also worth noting the multitype character of certain text categories. (p. 260)

This is a piece of advice worth noting.

THE BNC (BIBLIOGRAPHICAL) INDEX

The BNC Index spreadsheet I am about to describe was created as one solution to the previously mentioned problems and difficulties. It is similar to the plain text ones prepared by Adam Kilgarriff that I have benefited from and found rather useful.[14] 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.). Sebastian Hoffmann's files were useful too, in a complementary way, but these do not include (a) keywords and (b) the full bibliographical details of files. A third existing resource, the "bncfinder.dat" file that 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 BNC Index consolidates the kinds of information available in the above three resources, but, in addition, includes (a) BNC-supplied keywords (as entered in the file headers by the compilers); (b) COPAC keywords[15] for published non-fiction texts[16] (topic keywords entered by librarians); (c) full bibliographical details

(including title, date and publisher for written texts, and number of participants for spoken files); (d) 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; (e) 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). These extra notes are the result of random, manual checks: not all files have been subjected to such detailed analysis. 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 numeric code 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, and so forth, you need all the available information you can get. With the BNC 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. Changes and corrections to erroneous classifications were made both 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 BNC Index. Some of the genre categories are still being worked on, however, and may change in the final release of the Index.

Notes on the BNC Index

For spoken files, there are only eight relevant fields of information, giving the following self-explanatory details (abbreviations are explained in Table 6):[17]

File ID	Domain	Genre	Keywords	Word Total	Interaction Type	Mode	Bibliographical Details
FLX	S_cg_education	S_classroom	natural & pure science; chemistry	5,142	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 as follows:

File ID	Medium	Domain	Genre	Notes & Alternative Genres	COPAC Keywords	Keywords	Audience Age	Audience Sex	Audience Level
AE 7	book	W_nat_ science	W_non_ ac_nat_ science	Also W_non_ac_hu manities_arts	Biology - Philosophy	molecular genetics	adult	mixed	high

Bibliographical details	Total Words	Sampling	Circulation Status	Period Composed	Mode	Author Age	Author Sex	Author Type
The problems of biology. Maynard Smith, John. Oxford: OUP, 1989, pp. 9-109. 1686 s-units.	36,115	mid	M	1985-1994	W	60+ yrs	Male	Sole

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

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 **M**edium (this refers to the number of receivers of the text),[18] 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. 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)*,[19] and the text was composed in the period 1985-1994 *(Period Composed)*.[20] The *Bibliographical Details* field gives us the title of the text *(The Problems of Biology)*, its author, publisher, and so forth, 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 (the original bibliographical reference would have been "pp. ??"). However, some files did not have page breaks encoded at all, and thus their bibliographical references remain incomplete.

A list of all possible values for the closed-set fields (the keyword fields are open-ended) is given in Appendix B.

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 BNC 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, and so forth, you can find all such files by searching for "W_ac*" and specifying "high" for "audience level."[21] 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 very easily be done by using the *medium* field (i.e., by choosing either "book" or "periodical"), the sub-genres have not been given their own separate labels. Instead, end-users are encouraged to use available fields to create their own sub-classificatory permutations. The "genre" labels here are therefore meant to provide starting points, not a definitive taxonomy.

Table 5 shows the breakdown of the genre categories used in the BNC 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 hierarchical genre labels).

Table 5. Breakdown of BNC Genres in proposed classificatory scheme[22]

BNC SPOKEN	Super Genre	BNC WRITTEN	Super Genre
S_brdcast_discussn	Broadcast	W_ac_humanities_arts	Academic prose
S_brdcast_documentary		W_ac_medicine	
S_brdcast_news		W_ac_nat_science	
S_classroom		W_ac_polit_law_edu	
S_consult		W_ac_soc_science	
S_conv		W_ac_tech_engin	
S_courtroom		W_admin	
S_demonstratn		W_advert	
S_interview	Interviews	W_biography	
S_interview_oral_history		W_commerce	
S_lect_commerce	Lectures	W_email	Non-printed essays
S_lect_humanities_arts		W_essay_sch	
S_lect_nat_science		W_essay_univ	
S_lect_polit_law_edu		W_fict_drama	Fiction[23]
S_lect_soc_science		W_fict_poetry	
S_meeting		W_fict_prose	
S_parliament		W_hansard	
S_pub_debate		W_institut_doc	
S_sermon		W_instructional	
S_speech_scripted	Speeches	W_letters_personal	Letters
S_speech_unscripted		W_letters_prof	
S_sportslive		W_misc	
S_tutorial		W_news_script	
S_unclassified		W_newsp_brdsht_nat_arts	Broadsheet national newspapers
		W_ newsp_brdsht_nat _commerce	
		W_ newsp_brdsht_nat _editorial	
		W_ newsp_brdsht_nat _misc	
		W_ newsp_brdsht_nat _reportage	
		W_ newsp_brdsht_nat _science	
		W_ newsp_brdsht_nat _social	

	W_newsp_brdsht_nat_sports	
	W_newsp_other_arts	Regional & local newspapers
	W_newsp_other_commerce	
	W_newsp_other_report	
	W_newsp_other_science	
	W_newsp_other_social	
	W_newsp_other_sports	
	W_newsp_tabloid	Tabloid newspapers
	W_non_ac_humanities_arts	Non-academic prose (non-fiction)
	W_non_ac_medicine	
	W_non_ac_nat_science	
	W_non_ac_polit_law_edu	
	W_non_ac_soc_science	
	W_non_ac_tech_engin	
	W_pop_lore	
	W_religion	

It will be noted that aspects of this genre classification scheme mirror the ICE-GB corpus (see Table 5 for the ICE-GB categories), 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 broadsheet newspaper material together rather than into separate functional groups as in the ICE-GB (cf. "Reportage" and "Persuasive writing" in Table 5).

In some respects, the scheme also follows the Lancaster-Oslo/Bergen (LOB) corpus quite closely. This was done deliberately, to facilitate diachronic/comparative research.[24] For example, here is how the various subject disciplines are categorised in the LOB corpus and in the BNC Index:

Table 6. LOB Corpus Categories Broken Down into Component Disciplines

LOB (& BNC Index) Category	Subjects/Disciplines
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

One difference from the LOB corpus is that *economics* texts in the BNC 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, while *geography* is classed either as a social or natural science depending on the branch of geography. *Geology* has been classed as a natural science. 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.[25]

It should also be noted that some texts are 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 (see discussion which follows).

Some genres are deliberately broad because they can be easily sub-divided using other fields. For example, "institutional documents" includes government publications (including "low-brow'" informational booklets and leaflets/brochures), company annual reports, and university calendars and prospectuses. However, these texts can be fairly easily separated out using "Medium," "Audience level," or "Keywords."

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: (a) the "audience level (of difficulty)" estimated by the BNC compilers (coded in the file headers) (b) 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) (c) 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).

The spoken "lecture" genres in the Index refer only to university lectures. Thus, many "A"-level or non-university lectures are classified as "S_speech_unscripted." Similarly, "S_tutorial" refers only to university-level tutorials or classroom "seminars." Other non-tertiary-level or home tutorial sessions are classified under "S_classroom."

Genres labels are deliberately non-overlapping for spoken and written texts. For example, parliamentary speeches audio-transcribed by the BNC transcribers are labelled "S_parliament" for the spoken corpus, whereas the parallel, official/published version is labelled "W_hansard" for the written corpus. Also, for spoken texts, the "leftover" files (which do not really belong to any of the other spoken genres used in this scheme, e.g., baptism ceremony, auctions, air-traffic control discourse, etc.) are labelled as "S_unclassified," whereas leftover written files are labelled "W_misc."

As mentioned in the first part of this paper, 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 BNC 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; it helps to 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") are less well-defined than the written ones 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 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 BNC Index are advised to read the notes/documentation given here, and to be clear what the various domain and genre labels mean.[26] 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 "applied science." 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 BNC 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 makes 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.")

What About Library Classificatory Codes?

At this point, some people may be wondering if the classification systems used by libraries might be of use in helping us determine the proper genre labels. Atkins et al. (1992, p. 8) note in their discussion of the corpus attribute *topic* that "It is necessary to draw up a list of major topics and subtopics in the

literature. Library science provides a number of approaches to topic classification." This is an area that is beyond my expertise and the scope of this article, but I will make a few brief comments here.[27]

Several library classification/cataloguing systems are in use all over the world. They are all principally about subject areas (or topic) rather than about genre, although the two are, of course, related in many cases. A familiar scheme, the Dewey Decimal Classification system, is shown in Table 7.

Table 7. Dewey Decimal Classification System

Classmark	[Broad Area] & Subject Areas
Class 0	[GENERALITIES] Generalities; Catalogues; Newspapers; Computing
Class 1	[PHILOSOPHY & PSYCHOLOGY] Philosophy; Psychology; The Mind;
Class 2	[RELIGION] Religion
Class 3	[SOCIAL SCIENCES] Social Sciences; Law; Government; Society; Commerce; Education;
Class 4	[LANGUAGE] Linguistics; Scientific Study of Language
Class 5	[NATURAL SCIENCES & MATHEMATICS] Pure Sciences; Mathematics; Physics; Chemistry; Biology;
Class 6	[TECHNOLOGY (APPLIED SCIENCES)] Applied Sciences; Engineering; Medicine; Manufacturing;
Class 7	[THE ARTS] The Arts; (Music, Drama etc.) Recreations; Hobbies;
Class 8	[LITERATURE & RHETORIC] Literature
Class 9	[GEOGRAPHY & HISTORY] Geography; History; Information about localities

In addition to the Classmark, however, library materials are also given keywords which generally consist of Library of Congress subject headings (usually related to *topic[s]*). These are very useful when it comes to finding out what a text is about (or, in the case of fiction texts, what a text is).[28] In the case of literary texts, actual genre labels are sometimes given as keywords, and a frighteningly large number of sub-genres have been identified by the British Library cataloguers. These may prove useful to those who desire detailed sub-genre information on literary texts. A few examples will suffice here: Adventure stories, Detective and mystery stories, Picaresque literature, Robinsonades, Romantic suspense novels, Sea stories, Spy stories, Thrillers, Allegories, Didactic fiction, Fables, Parables, Alternative histories, Dystopias, Bildungsromane, Arthurian romances, Autobiographical fiction, Historical fiction, Satire, Christmas stories, Medical novels, Folklore, Domestic fiction, Ghost stories, Horror tales, Magic realism, Occult fiction, Feminist fiction, and Tall tales.

In addition to these fascinatingly categorised sub-genres,[29] the library also includes "form headings," which are meant to "define a type of fiction in terms of specific presentation, provenance, intended audience, form of publication."[30] Examples include Young adult fiction, Children's stories, Readers (Elementary), Plot-your-own stories, Diary fiction, Epistolary fiction, Movie novels, Scented books, Glow-in-the-dark books, Toy and movable books, Graphic novels, Radio and television novels, Sound effects books, Musical books, and Upside-down books.

As can be seen, therefore, library catalogues are a potentially valuable source of information as far as the *genre* classification of fiction texts and the identification of subject *topics* in non-fiction texts are concerned. Such information was, in fact, used in the process of creating the BNC Index, during the manual stage of checking and correcting the initial genre classifications I had made.

Using the BNC Index

The BNC 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).[31] 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, texts labelled "S_brdcast_discussion" also contain news reportage (in between the broadcast talk shows/programmes). This is unavoidable, since some BNC files combine genres and sub-genres and 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 which are 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*®, has been developed by Antonio Moreno Ortiz using the information contained in my spreadsheet.[32] A web-based facility, *BNC Web Indexer*, is also being developed at Lancaster, which does essentially the same thing.[33] 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*, 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,[34] 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.

It is beyond the scope of the present article to give more practical instructions or examples on how to use the BNC Index spreadsheet or the *Indexer* programs. Users will, in any case, surely find their own favourite ways of doing things, or may visit the relevant web sites for further information.

THE USES OF *GENRE*

In this paper, I have examined the different usages of the terms *genre, text type, register, domain, style,* and so forth. Which of these concepts is most useful for researchers, or for teachers to use in the context of classroom concordancing? I suggest that it is fruitful to start by looking at genres (categories of texts), and end up by generalising (through induction) about the existence of registers (linguistic characteristics) or even "text types" in Biber's sense (categories of texts empirically based on linguistic characteristics). The work by Carne (1996), Cope & Kalantzis (1993), Flowerdew (1993), Hopkins & Dudley-Evans (1988), Hyland (1996), Lee (in press), McCarthy (1998a, 1998b), Thompson (in press), and Tribble (1998, 2000), to name but a few, show how a genre-based approach to analysing texts can yield interesting linguistic insights and may be pedagogically rewarding as well. Thompson's paper, for example, shows how genre-based cross-linguistic analyses of travel brochures and job advertisements can reveal subtle, linguistically-coded differences in culture and point of view. Such genre analyses of relatively small, focussed and manageable sets of texts are now possible with the help of the BNC Index, opening up a rich resource for all kinds of learning and research activities. By searching for keywords in the various database fields, teachers and researchers can now quickly find even such rare sub-genres as

postcards, lecture notes, shopping lists and school essays ("rare" in the sense that they were not included in previous-generation general corpora and are hard to get hold of in machine-readable format even nowadays).

The personal BNC Index project described here is an attempt at classifying the corpus texts into genres or super-genres, and putting this and other types of information about the texts into a single, information-rich, user-friendly resource. This Index may be used to navigate through the mass of texts available. Users can then see at once how many texts there are that match certain criteria, and the total number of words they constitute. In this way, sub-corpora can then be easily created for specialised research or teaching/learning activities (e.g., it is now easy to retrieve BNC texts for ESP lessons to do with law, medicine, physics, engineering, computing, etc.).

Ultimately, one would wish that a deeper understanding of genres (their forms, structures, patterns) would be a "transformative" exercise for all investigators. As Cranny-Francis (1993) says,

> Genre is a category which enables the individual to construct critical texts; by manipulating genre conventions to produce texts which engender [critical analysis.] It also enables, therefore, the construction of a new, different consciousness …
>
> A concept of genre allows the critic or analyst to explore [the] complex relationships in which a text is involved, relationships which ultimately relate back to what a text means. This is because what a text says and how it says it cannot be separated; this is fundamental to our notion of genre. Because of this, genre provides the link between text and context; between the formal and semantic properties of texts; between the text and the intertextual, disciplinary and technological practices in which it is embedded. (pp. 111-113)

I hope that the disparate users and potential users of the BNC, whether researchers, teachers or students, will find the genre-enhanced BNC Index useful for all kinds of linguistic enquiry, and that some of the above transformative goals will be realised for them.

APPENDIX A

SPOKEN BNC Sampler: Missing or Unrepresentative Genres

- Consultations: medical (none)
- Consultations: legal (none)
- Classroom discourse (only 3 texts)
- Public debates (only 3 texts)
- Job interviews (none)
- Parliamentary debates (none)
- News broadcasts (none)
- Legal presentations (there are 2 legal cross-examinations, but no presentations, i.e., monologues)
- University lectures (none)
- Telephone conversations (no pure telephone conversations in the BNC as a whole)
- Sermons (only 1 text)
- Live sports discussions (none)
- TV/radio discussions (only 4 texts)
- TV documentaries (only 2 texts)

WRITTEN BNC Sampler: Missing or Unrepresentative Genres

- Academic prose: humanities (none)
- Academic prose: medicine (none)
- Academic prose: politics, law and education (only 2 texts on law, none on politics or education)
- Academic prose: natural sciences (nothing on chemistry, only 1 on biology & 3 on physics)
- Academic prose: social sciences (nothing on the core subject areas of sociology or social work, nor on linguistics, which is arguably a social science, even though it is often treated as a humanities subject)
- Academic prose: technology & engineering (nothing on engineering)
- Administrative prose (only 1 text)
- Advertisements (none)
- Broadsheets: the only broadsheet material included consisted entirely of foreign news, and only from the Guardian.
- Broadsheets: sports news (none)
- Broadsheets: editorials and letters (none)
- Broadsheets: society/cultural news (none)
- Broadsheets: business & money news (none)
- Broadsheets: reviews (none)
- Biographies (none)
- E-mail discussions (none)
- Essays: university (only 1 text)
- Essays: school (none)
- Fiction: Drama (only 1 text)
- Fiction: Poetry (only 2 texts)
- Fiction: Prose (insufficient texts, and only 1 short story)
- Parliamentary proceedings/Hansard (none)
- Instructional texts (none)
- Personal letters (none)
- Professional letters (none)
- News scripts (only 1 radio sports news script)
- Non-academic: humanities (only 2 texts)
- Non-academic: medicine (none)
- Non-academic: pure sciences (none)
- Non-academic: social sciences (2 rather odd texts, and 1 which possibly could be non-academic)
- Non-academic pure science material (i.e. popularisations of science texts: there were none of these in the Sampler)
- News scripts (classified as 'written-to-be-spoken' in the main BNC. None included in the Sampler)
- Official documents (only 1 text)
- Tabloid newspapers (only *Today* and *East Anglian Daily Times*, the latter of which is not really a tabloid, but a regional newspaper)

APPENDIX B
Information Fields and Possible Values in the BNC Index (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 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, laws, 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, W_fict_poetry, W_fict_prose** (drama, poetry and novels), **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, W_letters_prof** (personal and 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), **W_newsp_other_commerce, W_newsp_other_report, W_newsp_other_science, W_newsp_other_social, 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)

NOTES

1. In contrast, Nuyts (1988) uses "text type" in a rather idiosyncratic way to mean "a variety of written text" (as opposed to "conversation type" for spoken texts). Many other people similarly use "text type" in a rather loose way to mean "register" or "genre."
2. EAGLES is the Expert Advisory Group on Language Engineering Standards, an initiative set up by the European Union to create common standards for research and development in speech and natural language processing. At present, most EAGLES documents take the form of preliminary guidelines from which it is hoped that standards will later emerge.
3. In Biber's (1989) article on text typology, the nature of his "internal criteria" are more clearly shown. His "text types" are groupings of texts based on statistical clustering procedures which make use of co-occurrence patterns of surface-level linguistic features.
4. Wikberg (1992, p. 248) calls these rhetorical types "discourse categories" (German *Texttyp),* as opposed to "text types" (German *Textsorte)* which is equivalent to what I am here calling genres.

5. The GeM project at Stirling University illustrates an interesting new usage of genre. As it says on their Web site, "The GeM project analyses expert knowledge of page design and layout to see how visual resources are used in the creation of documents, both printed and electronic. The genre of a page -- whether it's an encyclopaedia entry, a set of instructions, or a Web page -- plays a central role in determining what graphical devices are chosen and how they are employed …. The overall aim of the project is to deliver a model of genre [italics added], layout, and their relationship to communicative purpose for the purposes of automatic generation of possible layouts across a range of document types, paper and electronic."
6. This diagram is from Martin (in press), but a similar one may be found in Eggins & Martin (1997, p. 243).
7. On a more speculative note, we could perhaps borrow from the tagmemic/particle physics perspective and talk in terms of particles (registers), waves (styles) and fields (genres). (Mike Hoey, personal communication.)
8. Martin (1993, 121) uses the term "macro-genre" to mean roughly the same thing.
9. Also, face-to-face conversations do not, arguably, form a proper genre as such (cf. Swales, 1990). However, for many research purposes, they form a coherent, useful super-genre.
10. Perhaps "religion" could also be considered a very broad content or topic label (?). In any case, this exceptional category apparently came about due to the unique nature of the texts: the corpus compilers note that the texts could "embrace any of the stylistic characteristics of [several other LOB categories]," yet they all belonged together in some sense. All "committed religious writing" was therefore put together under "Religion" (cf. Johansson, Leech, & Goodluck, 1978, 16).
11. As the EAGLES (1996) authors say, where there is a division into "factual" (informative) vs. "fictional" (imaginative), then "to avoid controversy, religious works are given a separate category of their own" (p.8).
12. Available on the Web at ftp://ftp.itri.bton.ac.uk/pub/bnc/bib-dbase. Titles of files in this resource are truncated to the first 80 characters, which limits its usefulness for some purposes.
13. The quote also contains an example of the term text types being used in a non-technical/loose fashion to mean "types/varieties of text."
14. 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.
15. 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. Keywords were manually copied from the Web catalogue entries and put into a separate column in the BNC 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.
16. For an explanation of why only non-fiction works are given keywords, see note 28.
17. Note that for the demographic files (conversations) the Keywords field is empty for almost all the files.
18. 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.

19. 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).
20. This follows the new 4-way classification scheme employed in the BNC World Edition: alltim0 (---[unclassified]); alltim1 (1960-1974); alltim2 (1975-1984); alltim3 (1985-1994).
21. Using "audience level=high" will roughly filter out introductory textbooks and texts written for both an academic and a more general audience.
22. Some of the genre names in the actual spreadsheet are further abbreviated for practical reasons.
23. Note that, in addition, there are four BNC files (EUY, HD6, KA2, KAV) which contain a roughly even mix of poetry and prose. These have been placed under the "W_misc" genre.
24. 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 using these corpora.
25. People who disagree with these classifications may use the "Keywords" and "Title" fields to find the relevant files and re-classify them as desired.
26. The domain labels in the BNC 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.
27. The British Library Web site (http://www.bl.uk) offers some detailed information & links.
28. A British Library "Fiction Indexing Policy" document states, "When indexing non-fiction it is right to attempt to express what the work as a whole is about, since it is usual for non-fiction to focus on one or more specific topics. By contrast, a work of fiction is rarely 'about' a topic at all. Instead, most works of fiction contain within them subjects as themes or settings. What they are 'about' is conveyed in the story as a whole. It is only themes, settings and characters which can be picked out easily by means of subject headings" (see http://www.bl.uk/services/bsds/nbs/marc/655polc.html).
29. As the EAGLES (1996) authors further point out, there are "alarming possibilities of double classification [i.e., mixed genres] -- spy thriller, historical romance, etc."
30. From the document at http://www.bl.uk/services/bsds/nbs/marc/655list2.html, which also gives a full listing of the literary sub-genres identified by the British Library.
31. The BNC Index spreadsheet, when ready, will be distributed initially at http://members.xoom.com/davidlee00/corpus_resources.htm. Suggestions for hosting on other sites are welcome.
32. Available at http://personal5.iddeo.es/tone/BNCIndexer. It is priced at 50 Euros for either an individual or institutional licence (up to 15 users).
33. BNC Web Indexer is the result of a collaboration between Paul Rayson (UCREL, Lancaster University) and myself. The URL will be announced on the CORPORA and CLLT (Corpus Linguistics and Language Teaching) mailing lists when available.

34. 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.

ABOUT THE AUTHOR

David YW Lee recently completed his doctoral studies at Lancaster University and is currently a visiting researcher and part-time tutor there. His PhD research involved applying Douglas Biber's multidimensional (MD) methodology to fresh spoken and written data from the British National Corpus (BNC) and a consequent critique of that factor-analysis-based methodology. At present, he is working on publishing his findings as a book, and is writing various articles for journals.

E-mail: david_lee00@hotmail.com

REFERENCES

Atkins, S., Clear, J., & Ostler, N. (1992). Corpus Design Criteria. *Journal of Literary and Linguistic Computing, 7*(1), 1-16.

Bhatia, V. (1993). *Analysing genre: Language use in professional settings.* London: Longman.

Biber, D. (1988). *Variation across speech and writing.* Cambridge, UK: Cambridge University Press.

Biber, D. (1989). A typology of English texts. *Linguistics, 27*(1), 3-43.

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing, 8*(4), 243-257

Biber, D. (1994). An analytical framework for register studies. In D. Biber & E. Finegan (Eds.), *Sociolinguistic perspectives on register* (pp. 31-56). New York: Oxford University Press.

Biber, D. (1995). *Dimensions of register variation: A cross-linguistic comparison.* Cambridge, UK: Cambridge University Press.

Biber, D. & Finegan, E. (1986). An initial typology of text-types. In J. Aarts & W. Meijs (Eds.), *Corpus linguistics II* (pp. 19-46). Amsterdam: Rodopi.

Biber, D., & Finegan, E. (1989). Drift and the evolution of English style: A history of three genres. *Language, 65,* 487-517.

Burnard, L. (Ed.). (1995, April 25). *The British national corpus users reference guide* (SGML version, First release with version 1.0 of BNC). Oxford, UK: Oxford University Computing Services.

Carne, C. (1996). Corpora, genre analysis and dissertation writing: An evaluation of the potential of corpus-based techniques in the study of academic writing. In S. Botley, J. Glass, T. McEnery, & A. Wilson (Eds.), *Proceedings of teaching and language corpora 1996, UCREL Technical Papers Vol. 9* (pp. 127-137). Lancaster, UK: Lancaster University.

Cope, B., & Kalantzis, M. (1993). Introduction: How a genre approach to literacy can transform the way writing is taught. In B. Cope & M. Kalantzis (Eds.), *The powers of literacy: A genre approach to teaching writing* (pp. 1-21). London: Falmer Press.

Cope, B., & Kalantzis, M. (Eds.). (1993). *The powers of literacy: A genre approach to teaching writing.* London: Falmer Press.

Couture, B. (1986). Effective ideation in written text: A functional approach to clarity and exigence. In B. Couture (Ed.), *Functional approaches to writing: Research perspectives* (pp. 69-91). Norwood, NJ: Ablex.

Cranny-Francis, A. (1993). Genre and gender: Feminist subversion of genre fiction and its implications for cultural literacy. In B. Cope & M. Kalantzis (Eds.), *The powers of literacy: A genre approach to teaching writing* (pp. 116-136). London: Falmer Press.

Crombie, W. (1985). *Discourse and language learning: A relational approach to syllabus design.* Oxford, UK: Oxford University Press.

Crystal, D., & Davy, D. (1969). *Investigating English style.* London: Longman.

Crystal, D. (1991). *A dictionary of linguistics and phonetics.* Oxford, UK: Basil Blackwell.

Expert Advisory Group on Language Engineering Standards. (1996, June). *Preliminary recommendations on text typology.* EAGLES Document EAG-TCWG-TTYP/P. [Available at http://www.ilc.pi.cnr.it/EAGLES96/texttyp/texttyp.html]

Eggins, S., & Martin, J. R.. (1997). Genres and registers of discourse. In T. van Dijk, (Ed.), *Discourse as structure and process* (pp. 230-56). London: Sage.

Faigley, L., & Meyer, P. (1983). Rhetorical theory and readers' classifications of text types. *Text, 3,* 305-325.

Fairclough, N. (1992). *Discourse and social change*. Cambridge, UK: Polity Press.

Fairclough, N. (2000). *New labour, new language?* London: Routledge.

Ferguson, C. (1994). Dialect, register and genre: Working assumptions about conventionalization. In D. Biber & E. Finegan (Eds.), *Sociolinguistic perspectives on register* (pp. 15-30). New York: Oxford University Press.

Finegan, E., & Biber, D. (1994). Register and social dialect variation: An integrated approach. In D. Biber & E. Finegan (Eds.), *Sociolinguistic perspectives on register* (pp. 315-347). New York: Oxford University Press.

Flowerdew, J. (1993). An educational or process approach to the teaching of professional genres. *ELTJ, 47,* 4305-4316.

Grishman, R., & Kittredge, R. (Eds.). (1986). *Analyzing language in restricted domains: Sublanguage description and procesing.* Hillsdale, NJ: Lawrence Erlbaum.

Halliday, M. A. K., & Hasan, R. (1985). *Language context and text: Aspects of language in a social-semiotic perspective*. Oxford, UK: Oxford University Press.

Hammond, J., Burns, A., Joyce, H., Brosnan, D., & Gerot, L. (1992). *English for social purposes: A handbook for teachers of adult literacy.* Sydney: National Centre for English Language Teaching and Research, Macquarie University.

Hoey, M. (1983). *On the surface of discourse.* London: Allen and Unwin.

Hoey, M. (1986). Clause relations and the writer's communicative task. In B. Couture (Ed.), *Functional approaches to writing: Research perspectives* (pp. 120-141). Norwood, NJ: Ablex.

Hopkins, A., & Dudley-Evans, T. (1988). A genre-based investigation of the discussion sections in articles and dissertations. *English for Specific Purposes, 7,* 113-121.

Hyland, K. (1996). Talking to the academy: Forms of hedging in scientific research articles. *Written Communication, 13*(2), 251-282.

Johansson, S., Leech, G., & Goodluck, H. (1978). *Manual of information to accompany the Lancaster-Oslo/Bergen corpus of British English, for use with digital computers.* Oslo: Department of English, University of Oslo.

Joos, M. (1961). *The five clocks.* New York: Harcourt Brace & World.

Kennedy, G. (1998). *An introduction to corpus linguistics.* London: Longman.

Kittredge, R., & Lehrberger, J. (Eds.). (1982). *Sublanguage: Studies of language in restricted semantic domains.* Berlin: Walter de Gruyter.

Kress, G. (1993). Genre as social process. In Cope, B., & Kalantzis, M. (Eds.), *The powers of literacy: A genre approach to teaching writing* (pp. 22-37). London: Falmer Press.

Kress, G., & Hodge, R. (1979). *Language as ideology.* London: Routledge & Kegan Paul.

Lee, David Y. W. (2000). Modelling variation in spoken and written language: The multi-dimensional approach revisited. Unpublished doctoral dissertation, Lancaster University.

Lee, David Y. W. (in press). Defining core vocabulary and tracking its distribution across spoken and written genres: Evidence of a gradience of variation from the British National Corpus. *Journal of English Linguistics.*

Martin, J. R. (in press). *Cohesion and texture.* Manuscript submitted for publication.

Martin, J.R. (1993). A contextual theory of language. In Cope, Bill & Mary Kalantzis (Eds.), *The Powers of Literacy: a genre approach to teaching writing* (pp. 116-136). London: Falmer Press.

McCarthy, M. (1998a). Taming the spoken language: Genre theory and pedagogy. *The Language Teacher, 22*(9). Retrieved June 20, 2000 from the World Wide Web: http://langue.hyper.chubu.ac.jp/jalt/pub/tlt/98/sep/mccarthy.html.

McCarthy, M. (1998b). *Spoken language and applied linguistics.* Cambridge, UK: Cambridge University Press.

Meyer, B. (1975). *The organisation of prose and its effects on recall.* New York. North Holland.

Nakamura, J. (1986). Classification of English texts by means of Hayashi's Quantification Method Type III. *Journal of Cultural and Social Science, 21,* 71-86.

Nakamura, J. (1987). Notes on the use of Hayashi's Quantification Method Type III for classifying English texts. *Journal of Cultural and Social Science, 22,* 127-145.

Nakamura, J. (1992). Hayashi's Quantification Method Type III: A tool for determining text typology in large corpora. An annex to a general report on annotation tools of the NERC Report. Unpublished manuscript.

Nakamura, J. (1993). Statistical methods and large corpora: A new tool for describing text types. In M. Baker, G. Francis, & E. Tognini-Bonelli (Eds.), *Text and technology: In honour of John Sinclair* (pp. 293-312). London: John Benjamins.

Nuyts, J. (1988). *IPrA survey of research in progress.* Wilrijk, Belgium: International Pragmatics Association.

Paltridge, B. (1995). Working with genre: A pragmatic perspective. *Journal of Pragmatics, 23,* 393-406.

Paltridge, B. (1996). Genre, text type, and, and the language classroom. *ELT Journal, 50*(3), 237-243.

Paltridge, B. (1997). *Genre, frames and writing in research settings.* Amsterdam: John Benjamins.

Phillips, M. A. (1983). Lexical macrostructure in science text. Unpublished doctoral dissertation, University of Birmingham, UK.

Rosch, E. (1973a). On the internal structure of perceptual and semantic categories. In T. E. Moore, (Ed.), *Cognitive development and the acquisition of language* (pp. 111-144). New York: Academic Press.

Rosch, E. (1973b). Natural categories. *Cognitive Psychology, 4,* 328-350.

Rosch, E. (1978). Principles of categorisation. In E. Rosch, & B. Lloyd (Eds.), *Cognition and categorisation.* Hillside, NJ: Lawrence Erlbaum.

Sampson, J. (1997). "Genre," "style" and "register". Sources of confusion? *Revue Belge de Philologie et d'Histoire, 75*(3), 699-708.

Steen, G. (1999). Genres of discourse and the definition of literature. *Discourse Processes, 28,* 109-120.

Stubbs, M. (1996). *Text and corpus analysis: Computer assisted studies of language and culture.* Oxford, UK: Blackwell.

Swales, J. (1990). *Genre analysis: English in academic and research settings.* Cambridge, UK: Cambridge University Press.

Taylor, J. R. (1989). *Linguistic categorisation: Prototypes in linguistic theory.* Oxford, UK: Clarendon.

Thompson, G. (in press). Corpus, comparison, culture: Doing the same things differently in different cultures. In M. Ghadessy, R. Roseberry, & A. Henry (Eds.), *The use of small corpora in language teaching.* Manuscript submitted for publication.

Tribble, C. (1998). *Writing difficult texts.* Unpublished doctoral dissertation, Lancaster University.

Tribble, C. (2000). Genres, keywords, teaching: towards a pedagogic account of the language of Project Proposals. In L. Burnard, & T. McEnery (Eds.), *Rethinking language pedagogy from a corpus perspective: Papers from the third international conference on teaching and language corpora* (Lodz Studies in Language; pp. 75-90). Hamburg: Peter Lang. Retrieved June 20, 2000 from the World Wide Web: http://ourworld.compuserve.com/homepages/Christopher_Tribble/Genre.htm.

van Dijk, T. (Ed.). (1985). *Handbook of discourse analysis.* London: Academic Press.

Wikberg, K. (1992). Discourse category and text type classification: procedural discourse in the Brown and the LOB corpora. In Leitner, Gerhard (Ed.), *New directions in English language corpora: Methodology, results, software developments* (pp. 247-261). Berlin: Mouton de Gruyter.