

TEXT CATEGORIES AND CORPUS USERS: A RESPONSE TO DAVID LEE

Guy Aston

University of Bologna, Italy

In designing any corpus, it is necessary to decide what types of texts to include, and how many of each type. (I use the term "text type" as a neutral one which does not imply any specific theoretical stance.) The *British National Corpus* (Burnard, 1995) made an initial division into written texts and spoken ones (i.e. transcripts of recordings), and within each of these macrocategories, employed further categorisations and subcategorisations. For the spoken component, a first distinction was between "demographic" (conversations: 153 texts) versus "context-governed" (speech recorded in particular types of setting: 757 texts), and the "context-governed" component was further divided according to the nature of the setting (educational/informative; business; public/institutional; leisure: from 131 to 262 texts in each), paralleled by a monologue/dialogue distinction (40%/60%). For the written component, two principal parallel categorisations were used: "domain" (i.e., subject matter, divided into nine classes, viz., *imaginative; arts; belief and thought; commerce; leisure; natural science; applied science; social science; world affairs*: from 146 to 527 texts in each) and "medium" (five classes, viz., *book; periodical; miscellaneous published; published; to-be-spoken*: from 35 to 1,414 texts in each). All figures refer to the *BNC World Edition* (2001).

Text categorisations, as Lee notes, are generally based on "external" criteria -- where/when the text was produced, by/for who, what it is about -- rather than "internal" ones based on its linguistic characteristics. The categorisations used in corpus design tend to be broad rather than delicate, since what corpus designers want to do is to enable users to generalize about and compare different categories. To generalize with any confidence, each category must contain a substantial *number* of different texts, so that no one text exerts an undue influence on that category (early corpora such as Brown and Lob, which were relatively small, got around this problem by including very short samples from a large number of texts); and each category must contain a wide *variety* of different texts, so that no one subcategory exerts an undue influence on that category as a whole (Biber, 1993): the greater the variance within a category, the more texts will be needed in order to document that variance. Thus, it may be decided to include roughly equal numbers of texts from different parts of the country, by authors of different sexes/ages or from different types of settings. Within the BNC "context-governed" component, for instance, the "educational/informative" category was designed to include lectures, talks, classroom interaction, and news commentaries, drawing these from different types of institutions in different areas and with a wide range of speakers and topics.

Since corpora cannot be infinite, the delicacy of the categorisations to be employed is largely determined by practical considerations. The BNC, which contains just over 4,000 texts, uses a framework which guarantees at least 100 texts in most principal categories. You may or may not like the categories chosen, but the corpus arguably allows you to generalize about these categories -- about spoken and written texts, the nine different domains of written texts, the four different domains of "context-governed" spoken texts, and so forth -- with reasonable certainty that findings will not be unduly biased by any particular text or any particular subcategory of texts. These categories are indicated in the headers to individual texts as attributes of the <catRef> element, using which it is possible to restrict queries to a particular category or combination of categories. A number of errors of categorisation in the first release of the BNC have been corrected in the *World Edition* (2001).

Users may, however, want to employ different categorisations from those employed by the corpus designers. David Lee provides one such categorisation, and the latest version of the SARA software (*SARA98*; Dodd, 2000) allows users to create their own subcorpora from the full BNC using his, or

other, categories (Aston, in press). Users should, however, be aware that such categories may be poorly represented in the corpus, both numerically and in terms of their variance. The more delicate the categorisation employed, the more likely it is that this will be the case (Sinclair, 1991) -- but even where a categorisation appears relatively broad, not all its members may be adequately documented. Thus Lee divides the BNC's *imaginative* written texts into *novels*, *poems*, and *drama*. However, there are only two texts in the BNC which fall into his *drama* category, so it would be pretty unwise to generalize about drama on their evidence. Why aren't there more? Some drama was included in the BNC in order to capture variance within the category of imaginative writing, but the quantity of drama is the result of decisions concerning the relative weight of drama in this category, just as the quantity of imaginative writing in the corpus is the result of decisions concerning the weight of imaginative writing in contemporary text production and reception as a whole. To include more drama would have either meant changing these design decisions or increasing the size of the corpus by an analogous factor.

All this means that if you want to generalize about contemporary British drama (or indeed about many of Lee's many other text categories), you would do much better to compile your own specialized corpus (though you may want to compare your findings with the BNC in order to see whether the features you identify are specific to the text-type in question). But you can't really complain about the BNC just because it doesn't contain more texts in a particular specialized category you happen to be interested in, whether this be e-mails, lectures, or business letters. That isn't what general mixed reference corpora are designed for, and you would clearly do better to start from a text archive instead, or from the Web.

But isn't a categorisation like Lee's what many users would like, and shouldn't the BNC have used such a categorisation to determine its composition? The main problem with Lee's approach, based on what he considers "prototypical" genres, is that it does not consider either the weight of these genres in the culture (in particular their frequency of reception and production), or the variance to be found within them. Lee appears to think that the BNC really ought to have provided representative samples for all 70 of his mutually-exclusive categories. But in order to include a minimum of, say, 50 texts in each category, either the corpus would have to have been very much larger, or else it would have had to weight these categories more or less equally ($70 \times 50 = 3,500$: the BNC contains just over 4,000 texts). Lee's three genres of imaginative writing (novels, poetry, and drama) would hardly seem to have the same frequency and variance within British culture, where much more fiction is read and published than poetry or drama, and, I suspect, of many more different kinds. So why should the corpus include the same amount of each?

Or take prayers. For some reason, prayers aren't one of Lee's genres, though I would have thought them as good a candidate for prototypical status as sermons, which are. There is only one text of prayers in the BNC, falling into the *to-be-spoken* written medium category (and into the *belief and thought* domain). The same *to-be-spoken* category, on the other hand, contains no fewer than 32 texts of television and radio news scripts. This disproportion seems fair enough when judged by production and reception standards -- news broadcasts play a much bigger part in British text production and reception than prayers do, alas. Yet, Lee's argument would suggest that they ought to have similar weighting, insofar as they have similar prototypical status (or else that the corpus should be much, much larger).

Lee's criticisms seem particularly unwarranted as far as the *BNC Sampler* (1999) is concerned (for the record, this contains no prayers, only one drama text, and only one news script). The *Sampler* -- which, like sampler music CDs, was designed to give a "taste" of the full BNC rather than to mirror its composition in detail -- consists of 184 texts for a total of roughly 2 million words, half speech and half writing. Lee complains that many of his categories are totally absent from the *Sampler*. But with this total number of texts, there is no way in which the *Sampler* could have adequately

documented 70 different categories while allowing reasonable generalizations at more macroscopic levels, such as speech versus writing. Would Lee really have wanted the number of university lectures on science in the *Sampler* to equal the number of casual conversations? Only, I think, if he were not interested in spoken texts in general, but particularly interested in science lectures, of which there would still not have been enough to say much about them.

A further problem with Lee's genre labels is that they may not match entire texts anyway. As he himself notes, virtually any single text may be analysed as composed of a number of different subtexts which can be assigned to different genres. For instance, there are 30 texts in the BNC consisting exclusively of poems, which Lee categorises as *W_fict_poetry*. However we find much more poetry occurring in texts belonging to other categories (as quotations, or when the hero of a novel breaks forth into song, etc.), 3,048 poems in 410 texts overall. Lee's categorisation is not going to be of much help to the user who wants to study poetry using the BNC. Rather than just those texts classed by Lee as poetry, s/he would be better advised to choose all those parts of the corpus texts which are tagged as <poem> elements in the markup (an easy task using SARA; Aston & Burnard, 1998).

With this last caveat in mind, where Lee does have a point is from what Gavioli (2001; Gavioli & Aston, 2001) calls an *example* rather than a *sample* perspective. Corpora like the BNC are designed to provide sample data from which to infer generalisations about the language as a whole, or about particular broad categories of texts, concerning frequencies of occurrence and co-occurrence (collocation, colligation, and so on). However it is also possible to use corpora -- at one's peril -- as text archives (Atkins, Clear, & Ostler, 1992) from which to retrieve examples of a particular text-type. If I am a teacher of religious education, and what I need for my lesson tomorrow is some prayers to use with my class -- why not look in the BNC? Since prayers are not a category used in the BNC text categorisation, to find candidate texts I will have to hope that either the text or its header (perhaps the text title, or its keywords) contains a form of the lemma *prayer* or a related word or phrase (perhaps *Amen*). A more detailed categorisation of the corpus texts, particularly one which uses prototypical "folk" genre labels, could be very useful as an aid to find examples of this kind.

This could also be a useful approach when we want to investigate a particular "user category" of texts. Not, I repeat, in order to generalize about that category, since the corpus cannot be relied upon to document it adequately, but in order to find examples from which to generate hypotheses. As mentioned earlier, there are 32 texts in the BNC containing radio or television news scripts (*W_news_script* in Lee's taxonomy). Given their limited number, and the fact that they come from a limited range of sources (two broadcasting stations), it would clearly be unwise to generalize from these to the genre of broadcast news scripts *tout court*. What they may provide, however, is a source of hypotheses about this genre -- hypotheses which must clearly be tested against a different corpus, one which has been constructed to comprise an adequately-sized sample of texts of this type, and which satisfactorily covers the variance within this category.

From an "example" perspective, the more descriptive categorizations that are provided within a corpus the better. For this reason, the incorporation of Lee's categories in the *BNC World Edition* (2001) is a very welcome development. For each text, his categorization forms the content of a <classCode> element in the header of the text (with the attribute *scheme="DLee"*), and using *SARA98* it is possible to restrict searches to one or more of his categories, and to define corresponding subcorpora -- subcorpora which can of course be adjusted if the user does not agree with Lee's attribution of particular texts to particular categories. I have, for instance, used a subcorpus of lectures from the BNC with a group of trainee conference interpreters who will need to work with academic monologue, selecting all those texts which Lee categorizes as lectures and then discarding two or three which seemed too informal and interactive for my purposes. There are nearly 50 lectures overall, on a wide range of topics and by a fair variety of lecturers, and it has proved a

useful collection from which to retrieve examples of particular discourse phenomena for teaching purposes and from which to generate hypotheses about the ways that lectures seem to work. Useful, that is, as long as you don't try to interpret it as a "representative sample" allowing reliable generalizations about lectures as a genre.

ABOUT THE AUTHOR

Guy Aston is Professor of English Linguistics in the School of Modern Languages for Interpreters and Translators at the University of Bologna, Italy. His main research interests concern the uses of corpora in language learning and in translation.

E-mail: guy@sslmit.unibo.it

REFERENCES

- Aston, G. (in press). The learner as corpus designer. In B. Kettemann (Ed.), *Teaching and language corpora 4* (provisional title). Amsterdam: Rodopi.
- Aston, G., & Burnard, L. (1998). *The BNC handbook: Exploring the British National Corpus with SARA*. Edinburgh, UK: Edinburgh University Press.
- Atkins, S., Clear, J., & Ostler, N. (1992). Corpus design criteria. *Literary and Linguistic Computing*, 7(1), 1-16.
- Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing*, 8(4), 243-257.
- The BNC Sampler*. (1999). Oxford, UK: Oxford University Computing Services.
- The BNC World Edition*. (2001). Oxford UK: Oxford University Computing Services.
- Burnard, L. (1995). *Users reference guide for the British National Corpus*. Oxford, UK: Oxford University Computing Services.
- Dodd, A. (2000). *SARA98*. Oxford, UK: Oxford University Computing Services.
- Gavioli, L. (2001). The learner as researcher: Introducing corpus concordancing in the classroom. In G. Aston (Ed.), *Learning with corpora* (pp. 108-137). Houston, TX: Athelstan.
- Gavioli, L. & Aston, G. (2001). Enriching reality: Language corpora in language pedagogy. *ELT Journal* 55(3), 238-246.
- Sinclair, J. McH. (1991). *Corpus, concordance, collocation*. Oxford, UK: Oxford University Press.