A2

For reasons we have already seen, the issue of size is far from easy to resolve in either theory or practice. In theory, as Halliday (1997) has suggested, a language is an infinitely large system, whereas the set of discourses in the language is always finite. In practice, every corpus has some cut-off range where — even if we can make an inventory of all the 'words' — we can never make an inventory of all the combinations that the language might allow. Corpus research confronts us with constant decisions about which regularities might encourage or discourage certain combinations, even when the data might look quite diverse, e.g., when the pejorative contextual cues of 'political instability' span 'enemy', 'foreign power', 'chaos', and 'anarchy', but also 'poverty' and 'reforms'.

As a corpus gets larger, we see that size improves the not just the quantity but the quality of the information we can get from the data. In his reports on COBUILD at 20 million, then 200 million, and most recently (as of June 1996) 323 million words, Sinclair has taken pains to refute the simple assumption that increases in size by no means merely follow a direct proportionality with the same data multiplied out, so that if an item appears once in a 1 million word corpus, it would appear 20 times in a 20 million word corpus and 200 times in a 200 million word corpus. Instead, we find numerous items that did not appear at all in smaller ones; we can make more informed judgements about relative frequency, e.g., when a small corpus shows two items appear only once each, whereas a larger corpus shows the one still only once and the other fifteen times; and an item which appeared only once in a small corpus may appear in several distinctive variants in a large one, e.g., 'indeterminate age' versus 'indeterminate years' in [212-14]. The proportionality assumption is no doubt derived from the further assumption, attractive to formalist but not to functionalist linguistics, that a 'language' is 'homogeneous in its linguistic characteristics' — just what corpus data soundly refute: 'there are important and systematic differences among text varieties at all linguistic levels', and 'global characterisations of "General English" should be regarded with caution' (Biber, Conrad, and Reppen 1994: 170, 179).

If very large corpora can reveal the 'heterogeneity' that prompted Saussurian formalist linguists to marginalise discourse data, the corpora can also offer us some means of defining it and determining how discourse participants normally manage it with fairly little time and energy. We can also explore the tendencies in various subdomains of a corpus or in specific sub-corpora. The major options pursued so far for sorting the domains or sub-corpora have been to apply either linguistic criteria, e.g., as 'text types' or 'language varieties', or else situational criteria, e.g., as 'registers' or 'professions'; not surprisingly, these two sets of criteria can produce quite divergent subdivisions and do not justify tidy borders separating them (cf. Biber 1989, 1994). Also, further differences keep emerging at greater degrees of detail, such as the subdivision of scientific or medical journal articles into 'methods' versus 'results and discussion' (cf. Biber and Finegan 1994).

My own proposal for sorting would be to co-ordinate the three dimensions of linguistic, cognitive, and social in order to construct multi-dimensional profiles of text types or discourse domains. We might begin with ones which, like medical journal articles, appear to be regulated by standardised conventions and move toward ones that appear less so, like family dinner conversations. How specific or general our criteria should be is a question to be tackled empirically as the research progresses, and to be co-ordinated with the applications we intend to support. Particularly if our findings are to be tapped in programmes for teaching English for Special (or Academic) Purposes, as Biber et al. (1994) in fact suggest, we could also inquire how far the prevailing conventions and strategies of the discourse serve purposes of inclusion or exclusion, and whether the degrees of specialisation are either necessary or productive (cf. Beaugrande 1997a, 1997b).

As we have seen, even a very large corpus of general English like the COBUILD can generate frequency statistics vulnerable to the periodic ‘ballooning’ effects caused by the shorter-range or longer-range preoccupations of public discourse with specific or fashionable topics. In July of 1994, when COBUILD’s Bank of English contained about 200 million words of running text, I found some striking ‘skews for news’: ‘revolutionary’ collocating 87 times with ‘Ethiopian’; ‘sex’ collocating 707 times with ‘Pistols’ and 63 times with ‘Madonna’, whose name occurred by itself some 2,516 times. Against these shorter-range preoccupations we can contrast some longer-range ones reflecting the voyeuristic if not indeed sadistic views our mass media seem to hold about what’s worth taking about: ‘death’ (31,013 occurrences), ‘dead’ (21,323), ‘died’ (22,467), ‘kill’ (51,746), ‘murder’ (18,383), ‘violence’ (19,226), ‘rape’ (5,890), ‘assault’ (4,055), ‘robbery’ (2,230), and ‘theft’ (1,970), as against a measly 661 occurrences of ‘kindness’ and just 10 of ‘human kindness’. The pet word of the modern age, ‘sex’, weighed in at 20,569 occurrences and collocated (aside from ‘Pistols’ and ‘Madonna’) predictably with ‘appeal’ (762) and ‘partner’ (120);

ominously with ‘offenders’ (247), ‘aids’ (117), ‘oral’ (203), ‘anal’ (108), ‘drugs’ (226), ‘violence’ (209), and ‘discrimination’ (209); and (perhaps?) benignly with ‘love’ (339) and ‘marriage’ (108) (I was in no mood to check out whether the ‘sex’ occurred with or without these last two collocates).

The shorter-range ballooning effects, provided they are distinctly lexical — the prospect of grammatical ones will be examined in just a moment — are fairly easy to spot; in 1997, who would suspect ‘Ethiopian’ as the principal collocate for ‘revolutionary’? And they could be offset by contrasting corpuses for different periods, e.g., subsequent decades, or by gradually accumulating one corpus over several decades. The longer-range ones, even if they are lexical, are more problematic and could be offset by shifting the bulk of the corpus away from mass media obsessed with violence and sex over toward everyday conversations in the home, the workplace, the evening party, and so on. Such is plainly desirable in theory; in practice, the labour and cost of putting them into a corpus are disheartening, and spoken data are still a small fraction of the total in, say, the COBUILD Bank of English or the British National Corpus. And of course we must wait and see how many everyday conversations are about ‘sex’ and ‘violence’ too.