New Resources, or Just Better Old Ones? The Holy Grail of Representativeness

Geoffrey Leech,Emeritus Professor,

Lancaster University

Abstract:

This paper runs counter to the majority of papers in this volume in focusing on the argument that, while welcoming opportunities to use new resources and methods, we should not neglect to improve and refine the resources and methods we already have.

The path of progress in corpus linguistics is strewn with unfinished business. Because no other realistic course is available, corpus linguists have understandably been following the path of practicality, pragmatism and opportunism. By and large, we have built up the resources and techniques of the present generation by taking advantage of what is already available and what can be relatively easily obtained. Our research efforts have consequently been limited and skewed by what resources we have been able to lay our hands on.

In this paper, I illustrate the skewing effect with reference to corpus design and composition, focusing on the desiderata of representativeness, 'balancedness' and comparability. After arguing that we need to give more consideration to these basic requirements, I briefly address the issue of representativity (a term used to mean 'the degree to which a corpus is representative') in relation to the use of the world-wide web as a source of corpus data.[1]

Comment [R1]: issues are also relevant to corpus building from www-material, web as corpus versus web for corpus compilation: discussion of this issue throughout the paper possible?

1. Introduction

In one sense corpus linguists appear to inhabit an expanding universe. The internet provides a virtually boundless resource for the methods of corpus linguistics. In addition, there is continuing growth in the number and extent of text archives and other text resources. If we consider corpora of the English language, one of the noticeable achievements has been the production of new historical textual resources,[2] so that gradually gaps in a mosaic of increasing coverage of historical varieties of the language are being filled in. This is greatly to be welcomed, obviously. Such are the increased opportunities for examining data of authentic usage in studying English that it may seem churlish to focus on what we lack, rather than on what new riches we can enjoy. On the other hand, there are still some weak spots in the coverage of natural language by existing corpora: notably in limitations in both quality and quantity of spoken language data, and in data from some of the newer electronic language media (e-mails, text messages, internet relay chat, and so forth).

2. Problems and challenges

One of our goals for the future should be to extend or refine existing resources: in other words, we need to strengthen the empirical foundations of corpus linguistics, not only in corpora but in the means to exploit them. There are many areas where corpus linguistics is not making appreciable progress. Strategies of stepwise refinement (for example, in corpus design and in POS-tagging) are known about, but are not activated. To take an example where research is skewed by what resources we can lay our hands on: Gaëtenelle Gilquin (2002) examined articles relating to grammar in the *International Journal of Corpus Linguistics* (IJCL) 1996-2001, and found that 68 per cent of these concentrated on word-based studies. Of the corpora used, 28 per cent were untagged 'raw' corpora, 43 per cent were POS-tagged corpora, and 29 per cent were parsed corpora. This suggests that the ways people use corpora have not caught up with the possibilities of sophisticated corpus analysis. The full potential of even limited annotation, that of part-of-speech tagging, has not been realised. Of course one can investigate English grammar using an untagged corpus, but this in general means that one can only investigate narrow areas of grammar where abstraction and generalization across lexical items are limited. Gilquin argued that we need a Holy Grail – the software capable of achieving a useful parsing of any corpus we want to investigate. So far an accurately working corpus parser has eluded us – although considerable human effort has been invested in the production of exceedingly useful parsed corpora, such as ICE-GB.

3. The Holy Grail of Representativeness

An even more basic issue at the foundations of corpus linguistics is: Have we been building the right kind of corpora?

It is generally accepted that one of the desiderata for a corpus is that it be REPRESENTATIVE, but in practice, this requirement has not been treated as seriously as it should be. A seminal article by Biber (1993) has frequently been cited, but no attempt (to my knowledge) has been made to implement Biber's plan for building a representative corpus. He came to the conclusion that the construction of such a corpus should "proceed in cycles: the original corpus design […] followed by collection of texts, followed by further empirical investigation of linguistic variation and revision of the design" (1993: 243). Although corpus linguists (including myself) often pay lip-service to representativeness, there has been relatively little productive debate on Biber's or anyone else's method of determining representativeness. However, one starkly negative contribution has been a paper by Váradi (2001), who dismisses the whole concept of representativeness as defined by Biber, and by implication claims that corpus linguistics is in a similar position to the emperor with no clothes. Much of the apologetics in favour of corpus linguistics stresses its immense advantages in providing a sound empirical base upon which to

formulate linguistic generalizations, explore variation, and test linguistic theories. But – looking at the matter with Váradi's sceptical eye – unless the claim that a corpus is representative can be substantiated, we cannot accept such findings. Without representativeness, whatever is found to be true of a corpus, is simply true of that corpus – and cannot be extended to anything else.

This is more serious than academic point-scoring. There is a crucial difference between claiming that such-and-such is the case in a corpus, and that the same such-and-such is the case in a language. By definition, a sample is representative if what we find for the sample also holds for the general population (Manning and Schütze, 1999: 119). Putting this in operational terms, 'representative' means that the study of a corpus (or combination of corpora) can stand proxy for the study of some entire language or variety of a language. It means that anyone carrying out a principled study on a representative corpus (regarded as a sample of a larger population, its textual universe) can extrapolate from the corpus to the whole universe of language use of which the corpus is a representative sample.[3] But as things stand at present, can we even claim a 'face validity' (to use a language testing term) for the representativeness of the corpora we work with?

This is, of course, taking a parole- or performance-based orientation towards language. For Chomsky and those taking his position, a corpus can only yield information about E-language (externalized language), and is therefore seen as irrelevant to the study of language per se, I-language (internalized language):

> Linguistics should be concerned with I-language and knowledge of I-language, that is with truths about the mind/brain, putting aside the irrelevant concept of E-language, however construed. (Chomsky, 1987: 45)

But for a corpus linguist, who specializes in the investigation of E-language, I take it that the goal of inquiry is to arrive, through the study of language in use, at a better understanding of some language, both in the sense of E-language and in the sense of I-language. The two are not in totally unconnected knowledge-domains, as Chomsky seems to assume. Rather, E-language is a crucial, indispensable manifestation of I-language. Yet the obvious point is that a corpus is a sample of E-language, not of I-language. The totality of a relevant textual universe of E-language is what is being sampled. For example, in the case of the *Lancaster-Oslo/Bergen* [LOB] *Corpus* (Johansson et al. 1978) the textual universe was the totality of published material produced by adult native speakers of British English published in the UK in 1961. This is a very large but finite textual universe, consisting of a finite number of texts of finite length. The same can be claimed about other corpora: the total textual universe of spoken utterances in the US in 1991 (say) is larger and more diffuse than the total textual universe of published texts of the same year. But it is still a finite (though mind-bogglingly large) set of utterances. It is true that lack of knowledge prevents us from enumerating the texts in this textual universe, and it is also true that the

linguistic domain of what is 'English' has some unclear boundaries. But this is a perfectly coherent and intelligible idea of what is being sampled, and I see no reason for Chomsky's claim that E-language is an 'epiphenomenon at best' (Chomsky 1986: 25), suffering from "complex and obscure socio-political historical and normative-teleological elements" (Chomsky 1991: 31).[4] Against this background, the claim that a corpus be representative of the textual universe of which it is a sample gains a sharper focus.

It is true that the textual universes associated with a modern language with a large number of native speakers, such as English, can be immense; but no more bafflingly immense than the universe of the material cosmos, about which physicists construct intelligible theories.[5]

4. What is a balanced corpus?

Another often-mentioned desideratum of a corpus is that it should be BALANCED, but there have been few attempts to explain what this requirement means. In my understanding, for a corpus to be balanced is an important aspect of what it means for a corpus to be representative. This 'balanced' quality has frequently been claimed for corpora such as the *Brown Corpus* or the *British National Corpus* [BNC] or ICE-GB, which have been carefully designed to provide sufficient samples of a wide and 'representative' range of text types. But balancedness is very difficult to demonstrate, even for such painstakingly constructed corpora. An obvious way forward is to say that a corpus is 'balanced' when the size of its subcorpora (representing particular genres or registers) is proportional to the relative frequency of occurrence of those genres in the language's textual universe as a whole. In other words, balancedness equates with proportionality. But no serious attempt was ever made to ensure that the genres in the Brown Corpus or the BNC were proportional in this sense. Váradi maintains that a corpus like the Brown Corpus is not representative in this sense, although its design was clearly intended to achieve some kind of proportionality, with some text categories being assigned many more text samples than others. He points out the immense difficulty of determining the proportional amount of text appropriate for just one text category, that of Humour, containing 9 of the 500 2,000-word texts in the corpus:

> For the BROWN corpus to qualify as a representative sample of the totality of written American English for 1963[6] for humorous writing, it would have to be established that humorous writing did make up 1.8% of all written texts created within that year in the US. (Váradi, 2001: 590)

It is instructive here to go back to the earliest discussions of corpus representativeness I am aware of, those that appeared in the volume edited by Bergenholtz and Schaeder (1979). Two contributors to that volume illuminated

the problem of representativeness in very different ways. Rieger (1979: 66) paradoxically claimed the pointlessness of achieving it:

> […] a random sample of the feature in question can only be designated representative when so much is known about the universe from which it comes that the formation of this sample is no longer needed.[7]

Bungarten (1979: 42-3) took a less negative stance, pointing out that even if we cannot achieve a representative corpus, there is a lesser degree of success worth achieving, what he usefully calls "an exemplary corpus":

> A corpus is exemplary, when its representativeness is not demonstrated, although less formal arguments, like evident coherence, linguistic judgements of competent researchers, specialist consensus, textual and pragmatic indicators, argue that the corpus may reasonably function as representative.[8]

Interestingly, it was in the same edited volume that Nelson Francis, chief begetter of the *Brown Corpus*, came up with a definition of a 'corpus' that included representativeness. A corpus, according to him, was 'a collection of texts assumed to be representative of a given language, dialect, or other subset of a language, to be used for linguistic analysis' (Francis 1979: 110). The tell-tale word here, of course, is 'assumed': there is nothing in the design of the Brown Corpus to guarantee representativeness. Instead, it seems that the Brown Corpus fits more snugly into the category Bungarten calls exemplary. Francis goes into some detail about the method of arriving at the composition of the Brown Corpus:

> […] we convened a conference of such corpus-wise scholars as Randolph Quirk, Philip Gove, and John B. Carroll. This group decided the size of the corpus (1,000,000 words), the number of texts (500, of 2,000 words each), the universe (material in English, by American writers, first printed in the United States in the calendar year 1961), the subdivisions (15 genres, 9 of 'informative prose' and 6 of 'imaginative prose') and by a fascinating process of individual vote and average consensus, how many samples from each genre (ranging from 6 in science fiction to 80 in learned and scientific).

Comment [R2]: on which page in Francis?

Unfortunately, the deliberations of these corpus-wise scholars have not come down to us: we do not know how far considerations of 'balance' led to their conclusion that 80 text samples were needed for the learned genre, and only 6 for that of science fiction. Although design of corpora has made considerable advances since that time, what makes a corpus 'balanced' or 'unbalanced' has remained obscure.

There is one rule of thumb that few are likely to dissent from. It is that in general, the larger a corpus is, and the more diverse it is in terms of genres and other language varieties, the more balanced and representative it will be.

However, perhaps we can do a little better than this. I would like to reconsider the value of proportionality in defining a balanced corpus. Biber (1993 – see also Biber et al 1998: 247) rejected proportionality, on the grounds that it would mean sampling of speakers and writers from the language community in proportion to their membership of demographic classes (e.g. by age, gender, socio-economic groupings, etc.), and this would lead to a highly *skewed* corpus, from the point of view of representing the whole range of linguistic variation, 90 per cent of the corpus consisting of conversation. Biber assumed that 90% of linguistic activity is conversational, and that conversation on the whole has relatively little variation compared with other varieties of language. He noted that other varieties of language would receive little representation (e.g. statutes, TV news) since only a tiny proportion of the language community is engaged in producing such texts.[9]

However, Biber elsewhere observes that there are three elements of language use that could enter into the sampling procedures. There are (a) the speakers and writers – the *initiators* of texts; (b) the hearers and readers – the *receivers* of texts; and (c) the *texts* themselves. I maintain that the representation of texts should be proportional not only to their initiators, but also to their receivers. After all, decoding as well as encoding is a linguistic activity. Thus a radio programme that is listened to by a million people should be given a much greater chance of being included in a representative corpus than a conversation between two people, with only one listener at any one time. I propose, therefore, that the basic unit to be counted in calculating the size of a given textual universe is not the text itself, but an initiator-text-receiver nexus, which we can call an ATOMIC COMMUNICATIVE EVENT (ACE). When a radio programme is listened to by a million people, there is only one text, but a million ACEs.

Since proportionality is widely considered to be the basis for representative sampling, Váradi (2001) criticizes Biber's (1993) decision to reject proportionality on the grounds of the estimation of greater 'importance' of certain genres (such as TV new broadcasts) in contrast to others (private conversations). Biber argues:

> It would […] be difficult to stratify a demographic corpus in such a way that it would insure representativeness of the range of text categories. Many of these categories are very important, however, in defining a culture. (Biber 1993: 245)

To which Váradi's riposte is:

> One of the fundamental aims of Corpus linguistics as I understand it is to show up language as it is actually attested in real life. However, Biber seems to argue that in designing a corpus one should apply a notion of

> importance that is derived from a definition of culture. … this throws the door wide open to subjective judgment in the compilation of the body of data that is expected to provide solid empirical evidence for language use. (Váradi, 2001: 592)

However, I would suggest that 'importance' does not have to be subjective. A conceptually simple way of measuring the importance of a text, for purposes of corpus building, is how many receivers it has. It is true that some corpus builders in the past have introduced evaluative criteria, judging, for example, a broadsheet newspaper (in the UK) to be more important or influential than a tabloid one; a novel which wins a national prize for literature to be more important than a pulp-fiction best-seller; or speakers belonging to socio-economic groups A and B to be more corpus-worthy than members of lower socio-economic groups D and E. However, this élitism is entirely spurious in a corpus intended for linguistic analysis.[10] In contrast, the criterion of size of readership/audience is free of evaluative bias. One of its results, no doubt unpalatable to corpus-builders with a sense of taste, is that tabloid newspapers are more likely to be included in a representative corpus than so-called quality or broadsheet newspapers. But this is something one has to put up with in the interests of representativeness.

It will not have escaped notice that the notion of an ACE as the basic unit of a textual universe, hence of sampling from a textual universe, is largely impractical. For the majority of samples we might want to include in a balanced corpus, we just have no way of knowing the number of texts, let alone the number of ACEs, in the relevant textual universe. The composition of the LOB Corpus was a particularly favourable case.[11] It was possible to use bibliographical sources to arrive at a relatively complete list of publications in the UK during the year 1961 to be used as a sampling frame. But no sampling on the basis of readership was attempted, and could not have been attempted for the corpus as a whole. It is true that some of the readership figures are relatively easy to come by – for example, the circulation of national newspapers – but for the large majority of publications, they are not. For most books, if we knew the number of copies purchased, we would be able to estimate the number of readers. But in general such information is not publicly available. A valuable source of information in the UK (and there are similar sources in other countries) is provided by the PLR (Public Lending Right) organization, which samples books borrowed from public libraries. But on the whole, the difficulties of determining the size of the textual universe and its sub-universes from which a corpus is to be sampled are formidable.

One additional difficulty is the variable length of texts. Text lengths, as well as text readerships, would have to be determined in order to calculate the likelihood that a sample of a given text should be included in a sample corpus. Thus a tabloid newspaper such as *The Sun* (in the UK) contains fewer words per issue than a broadsheet newspaper such as *The Independent*. This should give *The Independent* greater sampling privilege which would partially offset the smaller circulation of that paper.

Comment [R3]: This is actually an advantage of the www-approach, as access in something that is recorded on many web pages! Question: crawled-access also recorded?

It is reasonable to ask: Is there any point in pursuing the goal of a balanced corpus, where 'balanced' is understood to mean 'ACE-proportional' as it has been explained here? I will defend this concept of ACE-proportionality, while recognizing that it is a Holy Grail even more unattainable that Gilquin's working corpus parser. My arguments are these. First, the fact that something cannot be precisely specified or calculated does not detract from its actuality as something worth aiming at. Secondly, there are ways of mitigating the difficulty. (a) It is possible to estimate text usage, even where one cannot determine the quantity absolutely. (b) It makes sense to consider representativeness and balancedness in terms of a scale of approximation to the ideal. (c) Above all, representativeness (or, as I will prefer to call it, representativity) is a scalar phenomenon. Even if the absolute goal of representativeness is not attainable in practical circumstances, we can take steps to approach closer to this goal on a scale of representativity.[12] For example, the impossible calculations I referred to above can be estimated through the judgement of the corpus compilers combined with whatever objective measures may be available.[13]

In practice this is how people appear to have designed 'balanced' corpora in the past. The term 'judgement' here refers to the ability professionally competent members of a speech community seem to have in recognizing the relative prevalence of different genres, just as they may recognize their prevalent linguistic features. The 'corpus-wise' linguists who arrived at the composition of the Brown Corpus no doubt used this kind of judgement. A low degree of representativity, corresponding to Bungarten's 'exemplary corpus', can be attained by such informal means. A higher degree of representativity may be attained by using EXTERNAL (sociocultural)[14] criteria as formalized in a systematic typology of genres, as proposed by Biber (1989) among others. The aim here is to ensure that the widest practicable range of text categories within the textual universe is sampled. But we should perhaps emphasise the desirability of both breadth and depth in the typology: not only must the range be broad enough to include all genres at a primary level of classification, but the granularity of the typology must be sufficient to ensure that sampling includes delicate subcategories.

At a higher level still, representativity can be enhanced by a concerted effort to improve the proportionality of samples. This is, however, where I take a different route from Biber (1993: 248-55), who, having rejected proportionality, pursues a quantitative INTERNAL analysis of genres according to their linguistic characteristics. His plan is to carry out a multidimensional frequency analysis of register variation, and to develop a corpus which is representative of the full range of linguistic variation that occurs in the textual universe. Analysis of variation can reveal those registers where the corpus gives insufficient evidence of variation, and needs to be supplemented by additional textual material (longer textual samples, or more samples). By a cyclic research programme, the corpus can be gradually enlarged and modified until all variation in language use is sufficiently represented.

Comment [R4]: cross-reference to Biber, this volume?

Perhaps Biber's method is just another way of achieving balance. It will mean that language varieties are to be represented in the corpus in proportion to their heterogeneity, rather than in proportion to their prevalence of use in the whole textual universe. Arguably, this is not representativeness, but another corpus desideratum: heterogeneity. It is a different way of drawing the map of the varieties of usage. But the goal is similar: that once the map has been drawn, and the parameters of variation confirmed, the results of a corpus analysis can be extrapolated to data outside the corpus, and ultimately to the whole universe of language use.

How far is Biber's goal of an optimally heterogeneous corpus comparable to the ACE-proportionality theory of representativeness? It could have different results: for example, poets experiment endlessly with language, and the poetry genre is likely to show immense heterogeneity. But poetry might not score particularly highly in terms of volume of usage: poetry books do not tend to have a wide readership, nor poetry magazines a high circulation. So this might lead to a relatively low representation of poetry in a corpus modelled on ACE-proportionality, whereas it would lead to a high representation according to the heterogeneity.

Comment [R5]: principle/criterion?

Biber's method, like the ACE-proportionality, is extremely difficult to implement. One of the difficulties is that the size of text samples depends on the amount of text required to manifest a stable pattern of variation. With frequent grammatical characteristics, small text samples of 1,000 words are sufficient; however, as Biber admits, some linguistic features, such as a *that*-clause functioning as subject, are rare, and for these, much larger text samples would be needed. More dauntingly, if one considers collocations, lexico-grammatical combinations, granularity of linguistic classifications in grammar, in phonology, in semantics, etc., the number of linguistic features that might enter into a thorough study of variation is vast and open-ended. Some of these features would be very rare. The size of text samples needed would vary according to the linguistic feature under consideration, and for some rare features enormous text samples would be needed. There would be no 'fits all sizes' corpus. Biber et al (1998: 250) understandably comment that a great deal of work needs to be done in improving corpus design along these lines, and other difficulties, such as those of speech transcription, copyright clearance, or time and financial constraints, mean that compromises have to be made.

5. Comparable corpora

A third yardstick for successful corpus building is the construction of COMPARABLE CORPORA (also sometimes called 'matching' corpora): a set of two or more corpora whose design differs, as far as possible, in terms of only one parameter: the temporal or regional provenance of the textual universe from which the corpus is sampled. Thus, if comparability is achieved, one is entitled to assume that a significant contrast between one comparable corpus and another in

terms of linguistic frequency is likely to be due to the variability between the two corpora – of time or region – rather than variability within one corpus or within the other. The original example of comparability was that of the Brown and LOB corpora, which were intended to match one another in all respects apart from that of the country of origin (the USA versus the UK).

The requirement of comparability depends at least partly on that of representativity: comparable corpora permit precise comparisons between two varieties or states of a language, but only if the corpora are reasonably representative of their respective varieties. One might add, too, that comparability, like representativity, can be conceptualized as a scale, rather than as a goal to be achieved 100 per cent. The design profiles of the Brown and LOB corpora differed rather slightly, but enough to cause some doubt about whether we had truly attained a comparison of like with like.[15]

As is well known, a number of comparable corpora have been built on the Brown model, including the Frown [Freiburg-Brown] and FLOB [Freiburg-LOB] Corpora which match Brown and LOB respectively in being American and British matching corpora on the Brown model, but sampled from texts published in 1991/1992, rather than 1961. Hence the four corpora Brown, LOB, Frown and FLOB are each comparable in two dimensions, dialectal and diachronic: between American and British English, and between 1961 and 1991/1992. Another well-known example of comparable corpora is the *International Corpus of English* [ICE], where a corpus model (with stratified sampling from both written and spoken English) has been instantiated in different regional subcorpora such as ICE East Africa, ICE Great Britain, ICE India, ICE New Zealand, ICE Phillipines, ICE Singapore – these are the six varieties so far publicly available.

While it makes sense to achieve success in both representativeness and comparability, there is a sense in which these two goals conflict: an attempt to achieve greater comparability may actually impede representativity and vice versa. Nicholas Smith and I have encountered this problem in a mild form while building a 'prequel' to the LOB and FLOB corpora: a corpus on the familiar Brown model but with texts sampled from the years 1931±3 (i.e. 1928-34). Our most immediate research objective was to compare grammatical frequencies between 1931±3 and 1961, and to see how far they would enable us to project further into the past the trends already observed in the differences between the 1961 and 1991 corpora. But we encountered a problem with the sampling.

Rather like the wave and particle theories of light, representativeness and comparability, though each has its own validity, are ultimately incompatible ways of looking at corpus design. As one nears to perfection in comparability, one meets with distortion in terms of representativeness, and vice versa. In the LOB sampling, books and periodicals were randomly sampled[16] within the pre-determined text categories, from comprehensive lists of publications from the year 1961 (using the *British National Bibliography Cumulative Index 1960-64* and *Willings Press Guide*). When Christian Mair's Freiburg team set about building a 1991 equivalent of LOB, they aimed to achieve a one-to-one match between individual 2,000-word text samples in LOB and FLOB. This meant

choosing, for example, from the same periodicals if these happened to be in print both in 1961 and in 1991. Random sampling would not have achieved such a close match, and so would have jeopardized the comparability of the two corpora. For example, the compilers of the 1991 corpus "deliberately excluded papers which are circulated in vast quantities without charge. Although they are a sign of the times, we ranked the comparability of LOB '91 to LOB '61 higher in priority than the possible alternative goal, viz. to create the accurate picture of the British printed press right now" (Sand and Siemund, 1992: 120). In other words, comparability was prioritized at the expense of synchronic representativity. The Lancaster team have decided to follow this precedent in compiling the (so far incomplete) Lancaster1931 corpus. If we had followed the procedures of LOB, we would have carried out random sampling which would, for instance, have resulted in provincial newspapers from different cities being included in the Press categories A-C, and possibly the addition of new styles of publication (such as free newspapers) which had no equivalent in LOB.

This brings me to a more fundamental challenge to comparability: GENRE EVOLUTION (discussed in Leech and Smith, 2005). It is increasingly being recognized that the genres on which stratified sampling of many corpora is based are themselves subject to change. New genres emerge; old genres decay (see Biber et al. 1998: 252). As a case in point, we had problems filling the slots in the *Lancaster1931 Corpus* for science fiction and sociology texts – two sub-genres that were emergent at that time. One can argue that when these sub-genres are given the same degree of prominence in the 1931 corpus as in the 1961 corpus (where they were given 6 and 5 two-thousand-word samples respectively), they are overrepresented. Moreover, even some of the so-called sociology texts available in 1931 were arguably of a different genre, following more in the tradition of humanistic and philosophical discourse than in that of the then fledgling discourse of social science. We had to consider sample texts case by case, but in general followed the principle of text-by-text matching with LOB as far as possible. This policy, if adopted for an earlier corpus sampling publications in (say) 1901 or 1871, would clearly confront the compiler with more severe problems of genre definition, leading to increasing sacrifice of comparability to representativeness or vice versa, as one moved further into the past. The problems described here of maintaining diachronic comparability also arise with synchronic comparability. An example is the rearrangement of fiction text categories in the Australian Corpus of English, another corpus on the Brown model, where two new categories, Historical Fiction and Women's Fiction, were introduced, compensating for a dearth of Australia-published fiction in other categories.

The above discussion of representativeness, balance and comparability might lead the reader to reject these concepts as being ill-defined, problematic, unattainable. My attitude is different from this. I have tried to show that these are important considerations, and even if we cannot achieve them 100 per cent, we should not abandon the attempt to define them and achieve them. We should aim at a gradual approximation to these goals, as crucial desiderata of corpus design. It is best to recognize that these goals are not all-or-nothing: there is a scale of

representativity, of balancedness, of comparability. We should seek to define realistically attainable positions on these scales, rather than to abandon them altogether.

6. Conclusion: Internet Implications

I will finally turn to the theme of this book, and attempt to show how the reflections above have a bearing on the issue of using the web as a corpus. First, consider representativeness. One idea is that the Web-as-corpus makes the notion of a representative corpus redundant. Potentially, the whole of the Web can be searched with a search engine, so a sample corpus is unnecessary: we have the whole textual universe at our disposal. However, it is clear that this ideal situation does not exist. A search engine like Google employs algorithms which are totally mysterious to the user. Google provides nothing like a complete search of the Web, and reports such as that by Jean Véronis (http://aixtal.blogspot.com/2005/02/web-googles-missing-pages-mystery.html) show how unstable and inconsistent are the counts that one gets from Google, at least at the present time. What we get is an enormous sample of the Web, but how representative it is remains a mystery. The consensus seems to be that frequency information obtained from Google is at present seriously misleading.

What must be excluded from the above judgement, of course, are well-defined custom-made corpora based on particular websites, such as the CNN corpus and the SPEA-Corpus introduced by Hoffmann and Hundt/Biewer in their respective chapters of this book.

A second question, with regard to representativeness, is: Can we see the Web (or the sample of it we access in searching with a Web browser) as somehow representative of English language use as a whole; or at least of written English language use? Can the proportional sense of a 'balanced corpus' be applied to it? It is true that the Web gives access to a very wide range of genres, some of them well-established in the written medium, such as academic writing and fiction writing; others newly-evolving genres closer to speech, such as blogs. However, it is also true that the Web by definition gives little or no access to private discourse, such as everyday conversation, telephone dialogues, and the like. Searching with a search engine provides no access to spoken or manuscript data. There are major areas seriously underrepresented, if they are represented at all. It is also likely that certain varieties, such as academic writing, are overrepresented. The multi-media and HTML format of webpages is also likely to exercise its own constraints and preferences in the use of language. The Web in English is its own textual universe, with huge overlaps with the non-electronic textual universe of the English language. It is a textual universe of unfathomed extent and variety, but it can in no way be considered a representative sample of language use in general.

Turning to the concept of comparability: it is obvious that the Web provides nothing like the exact comparability of text selection for different

Comment [R6]: no, they are not, cross-reference to Fletcher, this volume!

Comment [R7]: web address in footnote?

Comment [R8]: We changed the name of our corpus, SPEAC stands for *South Pacific and East-Asian Corpus.*

Comment [R9]: the very notion of 'privacy' is sometimes challenged by www-mediated communication, isn't it?

Comment [R10]: internal variability of English, English www representative of it (thinking along the lines of world-wide English)?

periods or different regions of the world. On the diachronic axis, it is even impossible to tell when a particular text or text extract was written; similarly, on the synchronic axis, knowledge of the provenance of a text is minimal. Whether the author was a native speaker, for example, is unknown. On the other hand, searching on the country codes in URLs can provide convincing gross frequency contrasts between national varieties, as in the case of *different from*, *different than*, and *different to* illustrated in Mair's contribution to this book. If we are interested in rough-and-ready rather than more precise frequency data, and observe sufficiently striking contrasts, the Web can offer revealing results, which can be confirmed by replication.

Comment [R11]: a very strong claim!

Even without such qualities as representativeness, a corpus retains the merit, in Váradi's terms, in showing up "language as it is actually attested in real life". In providing evidence for neologisms, new word usages, and collocations the Web wins out against other corpora because of its sheer size and because it is always being updated. Hence it is useful, and may have even become indispensable, for lexicography and for lexico-grammatical investigations. The absence of any linguistic annotation such as POS tagging means that grammatical and semantic investigations are limited, in the ways indicated by Gilquin (2002). They have to rely on searches based on orthographic lexical form, which is not to say they are unimportant. Perhaps the future will bring 'intelligent search engines' which consign this restriction to history. Meanwhile, while the internet is an added resource of immense potential, it does not remove the need to improve and update other textual resources, and does not render obsolete the corpus compiled according to design and systematic sampling.

Comment [R12]: cross-reference to Fletcher, Renouf et al., this volume)

1 I am grateful to Nick Smith for helpful discussions on some topics covered in this paper.

2 Just a few of the new historical corpora arising from recent work are the *Penn-Helsinki Parsed Corpus of Middle English* (PPCME2), the *Penn-Helsinki Parsed Corpus of Early Modern English* (PPCEME), extensions of the ARCHER corpus (*A Representative Corpus of Historical English Registers*), the *Corpus of Early English Correspondence Sampler* (CEECS), the *Corpus of Late 18th Century Prose*, the *Corpus of English Dialogue* (1560-1760), the *Corpus of Nineteenth Century English* (CONCE), and the *Zürich Corpus of English Newspapers* (ZEN). In addition, proprietary resources like *The Times Digital Archive*, the *OED quotations database*, and the Chadwyck-Healey literature collections provide further extensive and rich full-text resources for the history of English.

3 Of course not all corpora are samples. Some corpora contain the complete extant textual material belonging to a certain language or language variety. Examples are the *Corpus of Shakespeare's Works*, the *Corpus of Hellenistic Greek*, the *Corpus of twentieth-century newspapers in Basque* (see the UZEI *Systematic Compilation of Modern-Day Basque* – EEBS).

Particularly in the case of languages long dead, the corpora of data that have come down to us are the result of chance survivals, of course contain no spoken language, and are usually heavily biased towards certain periods, genres, and authors. Porter and O'Donnell (2003: 121) observe that for Hellenistic Greek, "in the 55 million words in the Thesaurus Linguae Graecae database, around 10 million of those words are by the fourth-fifth century writer John Chrysostom". In the case of such closed exhaustive corpora, the issue of representativeness clearly cannot be seriously addressed.

4 These two quotations from Chomsky are found in Váradi (2001: 587).

5 Footnote missing

6 The year '1963' here is presumably an error for '1961'.

7 "[...] eine Stichprobe hinsichtlich des betrachteten Merkmals nur dann als *repräsentativ* ausgezeichnet werden kann, wenn über die Grundgesamtheit, aus der sie stammt, so viel bekannt ist, daß es eben dieser Strichprobenbildung gar nicht mehr bedarf." (See Mark Sebba's webpage http://www.ling.lancs.ac.uk/staff/mark/cwbc/cwbcman.htm)

8 „Ein Korpus ist exemplarisch, wenn seine Repräsentativität nicht nachgewiesen ist, andererseits weniger formale Argumente, wie evidenter Zusammenhang, linguistische Urteile des kompetenten Forschers, fachlicher Konsensus, textuelle und pragmatische Indikatoren, für eine sinnvolle Vertreterfunktion des Korpus plädieren." (Also quoted on Mark Sebba's webpage.)

9 Footnote missing

10 Interestingly, the BNC falls foul of the accusation of élitism on two of these counts, although the intentions behind these decisions were not élitist. The numbers of speakers sampled for the demographic (conversational) subcorpus from the lowest socio-economic classes D and E were equal to those from classes A and B, although if they had been sampled in proportion to population, they would have been larger. This apparent 'élitist' deviation from proportionality was reportedly due to a difficulty data-collecting researchers experienced in persuading members of classes D and E to record their own conversations and to take part in the data collection. In the interests of economy, the easier way out was chosen: the samples from each class were equalized. Another 'élitist' deviation from proportionality was the higher representation of broadsheet ('quality') newspapers in the BNC than of tabloid newspapers. The reason for this was that permission could not be obtained from certain newspapers to include their material in the corpus.

Comment [R13]: 'the' deleted

11 The reason for choosing LOB as a particularly favourable case, rather than Brown, is that the Brown texts were sampled from local libraries in Providence, R.I., where Brown University is located, and therefore contains a representation of US publications in 1961 that is limited to the holdings of these libraries.

12 This view is taken by Mukherjee (2004: 114): "Absolute representativeness is an unattainable ideal, but specific procedures may help in getting closer to this goal […]".

13 The following is an afterthought, added after this paper had been completed and submitted. The criterion of ACE-proportionality can, without much simplification, be reduced to a criterion of receptive proportionality. The argument is as follows. The number of ACEs (communicative events defined in terms of a single addresser and a single addressee) represented by a text is the product of the number of addressers and the number of addressees of that text. However, in all canonical cases the number of addressers (whether in speech or in writing) is just one. (Cases of multiple addressers are, of course, found in choral speech and in co-authorship of written texts, but these cases are confined to rather special circumstances: by far the majority of published texts, for example, have a single author.) Hence the number of ACEs per text reduces, without much distortion, to the number of receivers of a text. Research on proportionality therefore reduces to language reception research, which can be conducted along the sociological lines, taking a demographic sample and investigating (by means of diaries, questionnaires, etc.) the amount of time spent in listening to different categories of speech and reading different categories of written text. In the design of the *Czech National Corpus* this kind of language reception research was employed to determine proportions of different genres of written text (Cermák 2003: 212). To undertake a fully-fledged research project of this kind as a prerequisite to compiling a balanced corpus would be rather expensive and time-consuming, but not beyond the bounds of possibility. It would also be valuable for other research domains, such as literacy research.

14 The terms 'external' and 'internal' here follow the usage of Sinclair (1996), for whom text classification can be based either on external 'sociocultural' or internal 'text linguistic' criteria. Biber (1993) made use of a similar distinction.

15 For example, the Western and Adventure Fiction category (N) in Brown contained many more Western Fiction texts than the LOB Corpus, as such works were rarely published in the UK.

16 Problems of copyright clearance meant that random sampling was not adhered to in all cases.

References

Bergenholtz, Henning and Schaeder, Burkhard (eds.) (1979) *Empirische Textwissenschaft: Ausbau und Auswertung von Text-Corpora*. Königstein: Scriptor.

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

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

Biber, Douglas, Susan Conrad and Randi Reppen (1998) *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Bungarten, Theo (1979), „Das Korpus als empirische Grundlage in der Linguistik und Literaturwissenschaft". In Bergenholtz and Schaeder, pp. 28-51.

Chomsky, Noam (1986), *Knowledge of Language: Its Nature, Origin and Use*. New York: Praeger.

Chomsky, Noam (1987), *Generative Grammar: Its Basis, Development and Prospects*. Kyoto: Kyoto University of Foreign Studies.

Chomsky, Noam (1991), "Linguistics and cognitive science: Problems and mysteries." In Kasher, A (ed.) *The Chomskyan Turn*. Oxford: Blackwell, pp.26-53.

Francis, W. Nelson (1979), "Problems of assembling and computerizing large corpora." In Bergenholtz and Schaeder, pp. 110-123.

Johansson, Stig, Geoffrey Leech and Helen Goodluck (1978), *Manual of Information to Accompany the Lancaster-Oslo/Bergen Corpus of British English, for Use with Digital Computers*, Department of English, University of Oslo.

Gilquin, Gaëtanelle (2002), "Automatic retrieval of syntactic structures: the quest for the Holy Grail", *IJCL*, 7:2, 183-214.

Leech, Geoffrey and Nicholas Smith (2005), "Extending the possibilities of corpus-based research on English in the twentieth century: a prequel to LOB and FLOB", *ICAME Journal*, 29: 83-98.

Manning, Christopher D. and Hinrich Schütze (1999), *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Mukherjee, Joybrato (2004), "The state of the art in corpus linguistics: three book-length perspectives", *English Language and Linguistics* 8.1: 103-119.

Porter, Stanley E. and O'Donnell, Matthew (2003), "Theoretical issues for corpus linguistics and the study of ancient languages". In Wilson, Andrew, Paul Rayson and Tony McEnery (eds.), *Corpus Linguistics by the Lune*, Frankfurt am Main: Lang, pp. 119-137.

Rieger, Burghard (1979), „Repräsentativität: von der Unangemessenheit eines Begriffs zur Kennzeichnung eines Problems linguistischer Korpusbildung." In Bergenholtz and Schaeder, pp. 52-70.

Sinclair, John (1996), *Preliminary Recommendations on Corpus Typology*. EAGLES Document EAG—TCWG—CTYP/P http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html

Váradi, Tamás (2001), "The linguistic relevance of corpus linguistics". In Paul Rayson, Andrew Wilson, Tony McEnery, Andrew Hardie and Shereen Khoja (eds.) *Proceedings of the Corpus Linguistics 2001 Conference*, Lancaster University: UCREL Technical Papers 13, pp.587-593.