Methodological Issues Regarding Corpus-based Analyses of Linguistic Variation

DOUGLAS BIBER
Northern Arizona University, USA

Abstract

Although corpus-based analyses of linguistic variation have provided fresh insights into previously intractable issues, several methodological criticisms have been raised about the overall design of text corpora and the validity of text 'genres' as a basis for analyses of variation. Unfortunately, most of these criticisms have been based on intuitive judgements rather than empirical investigation. The present study begins to correct this lack of evidence concerning these issues. It focuses on four particular methodological issues: (1) how long texts should be in order to reliably represent the distribution of linguistic features in particular text categories; (2) how many texts within each text category are required in order to reliably represent the linguistic characteristics of that category, and related questions concerning the validity of 'genre' categories; (3) how many texts are needed in a corpus to accurately identify the salient parameters of linguistic variation among texts; and (4) how much of a cross-section is required to identify and analyze the salient parameters of variation among texts. These issues are addressed through statistical investigation of the distribution of linguistic features across various sub-samples of the LOB and London-Lund corpora, in comparison to their distribution across the full corpora. The results indicate that existing corpora are adequate for many analyses of linguistic variation. In conclusion, the paper welcomes the future availability of larger and more representative corpora, but it also urges researchers to fully exploit existing corpora for ongoing investigations of linguistic variation.

1. Overview of the issues

A corpus-based approach to linguistic issues has become increasingly popular in the last several years. The use of text corpora in American linguistics actually dates back to the early part of this century, when anthropologists such as Boas and Sapir collected corpora of folk-tales from Native American languages as databases for subsequent linguistic analysis; and this tradition continues to the present among anthropological linguists and social dialectologists who study corpora of tape-recorded speech. More recently, several text corpora have become available on computers, enabling analyses of a different type and scope than previously possible. Within linguistics, there are three main research communities that exploit the resources of computer-based text corpora: lexicographers (e.g. Sinclair, 1987; Walker, 1989), researchers in speech and natural language processing (e.g. Garside *et al.*, 1987), and sociolinguists studying language variation (e.g. Coates, 1983; Altenberg, 1984; Tottie and Bäcklund, 1986). Corpus design considerations differ for each of these user groups, depending on the characteristic linguistic analyses required by their research goals; the present paper focuses on the needs of the last group mentioned above (sociolinguists and descriptive linguists studying language variation).

Although many studies have used a corpus-based approach to analyze linguistic variation, there are several methodological issues regarding this approach that have not been adequately addressed to date. Oostdijk (1988), in a critique of corpus-based approaches to linguistic variation, paints a dismal picture of previous research; the paper concludes that studies based on existing corpora have not, and *cannot*, make interesting contributions to the analysis of linguistic variation among texts. Previous research is summed up as follows: 'corpus-based studies of linguistic variation have so far failed to make a substantial contribution to the development of a descriptive theory of linguistic variation' (p. 19), primarily because 'the corpora that were used were not suited for studying linguistic variation' (p. 14). Existing corpora are claimed to be inadequate because 'corpora which have [been] compiled with the intention of representing a cross-section of a language are not suited for the study of linguistic variation since, in selecting a great many different samples, they neutralize any variety-specificity (p. 17); that is, existing corpora are inadequate because 'their samples are too small' (p. 19).

The major text corpora referred to by Oostdijk are the Brown, LOB, and London-Lund corpora. The Standard Corpus of Present-Day Edited American English, known as the Brown Corpus, contains one million words of written, edited American English published in 1961; the corpus comprises 500 text samples, each 2,000 words long, taken from fifteen different text categories (e.g. press reportage, editorials, academic prose, general fiction). The LOB corpus is a British English parallel to the Brown Corpus. The London-Lund Corpus is based on the spoken texts collected as part of the Survey of English Usage; it comprises eighty-seven text samples, each at least 5,000 words long, representing several major speech situations (e.g. face-to-face and telephone conversations, radio broadcasts, interviews, public speeches). All these corpora are described in detail in Oostdijk's paper (cf. Kucera and Francis, 1967; Svartvik and Quirk, 1980; Francis *et al.*, 1982; Johansson, 1982; Johansson and Hofland, 1989.

Whether or not one agrees with its bleak assessment, Oostdijk's paper raises several important methodological issues, including the overall claim that existing corpora are 'not suited' for the study of linguistic variation. Unfortunately, the paper does not provide the empirical data needed to address these issues. In the present paper, I investigate some of these issues through empirical analyses of the distribution of linguistic features in a corpus of texts. In particular, I consider the issues of: (1) how long texts should be in order to reliably

Correspondence: D. Biber, Department of English, Northern Arizona University, Flagstaff, AZ 86011-6032, USA.

represent the distribution of linguistic features in particular text categories; (2) how many texts within each text category are required in order to reliably represent the linguistic characteristics of that category, and the validity of 'genre' categories; (3) how many texts are needed in a corpus to accurately identify the salient parameters of linguistic variation among texts; and (4) how much of a cross-section is required to identify and analyze the salient parameters of variation among texts.

These questions are investigated within the framework of Biber (1988), which undertakes an overall analysis of the parameters of register variation in English. That study uses a set of computer programs and computer-based corpora to analyze the distribution of sixty-seven linguistic features among twenty-three spoken and written text genres. In the present paper, the database and analytical techniques from the earlier study are used to investigate the methodological issues listed above. (The parameters of variation identified in the earlier study are discussed further in Section 4.)

In one sense the single answer to all four of these methodological questions is: 'It depends'—it depends on the goals of the analysis and on the particular linguistic features and kinds of texts being analyzed. However, the findings reported here argue against the claim that existing computer-based corpora are unsuited for the study of linguistic variation; instead, they indicate that, with respect to the above questions, the corpora are adequate for many studies of linguistic variation across texts.

2. The issue of text length

One of the important issues in corpus design is text length. Texts in the Brown and LOB corpora are 2,000 words long, while texts in the London-Lund corpus are 5,000 words in length. In contrast, Oostdijk states that much longer text samples are required in order to analyze linguistic variation, suggesting 20,000 words as an adequate length (p. 20). There is thus considerable disagreement on this issue.

To investigate this issue, I analyzed certain aspects of the internal variation within texts by comparing pairs of 1,000 word samples taken from single texts in the LOB and London-Lund corpora. If large differences are found between the two 1,000 word samples, then we can conclude that 1,000-word samples do not adequately represent the overall linguistic characteristics of a text and that possibly 2,000-word and 5,000-word samples are also inadequate. If, on the other hand, the two text samples are similar linguistically, indicating that even a 1,000-word sample reliably measures the linguistic characteristics of a text, we can be fairly confident using the 2,000-word and 5,000-word samples of the LOB and London-Lund corpora.

In the case of written texts (from the LOB corpus), I divided the original texts in half and compared the two parts. In the case of spoken texts (from the London-Lund corpus), four 1,000-word samples were extracted from each original text, and these were then compared pairwise. To provide the strongest test of the spoken materials, the paired 1,000-word samples were not contiguous in the original text.

To provide a relatively broad database, ten linguistic features commonly used in variation studies were analyzed. These features were chosen from different functional and grammatical classes, since each class potentially represents a different statistical distribution across text categories (see Biber, 1988). The features are: first person pronouns, third person pronouns, contractions, past tense verbs, present tense verbs, prepositions, WH relative clauses, passive constructions (combining *by*-passives and agentless passives), and conditional subordinate clauses. Pronouns and contractions are relatively interactive and colloquial in communicative function; nouns and prepositions are used for integrating information into texts; relative clauses and conditional subordination represent types of structural elaboration; and passives are characteristic of scientific or technical styles. These features were also chosen to represent a wide range of frequency distributions in texts, as shown in Table 1, which presents their frequencies (per 1,000 words) in a corpus of 481 spoken and written texts (taken from Biber, 1988, 77–8). As Table 1 shows, the ten features differ considerably in their overall average frequency of occurrence and in their range of variation. Nouns and prepositions are extremely common; present tense markers are quite common; past tense, first person pronouns, and third person pronouns are all relatively common; contractions and passives are relatively rare; and WH relative clauses and conditional subordinators are quite rare. (In addition, these features are differentially distributed across different kinds of texts; see Biber, 1988, 246–69.) Comparison of these ten features across the 1,000-word text pairs thus represents several of the kinds of distributional patterns found in English.

The distributions of these linguistic features were analyzed in 110 1,000-word text samples (i.e. fifty-five pairs of samples), taken from seven text categories: conversations, broadcasts, speeches, official documents, academic prose, general fiction, and romance fiction. These categories represent a range of communicative situations in English, differing in purpose, topic, informational focus, mode, interactiveness, formality, and production circumstances; again, the goal was to represent a broad range of frequency distributions.

Reliability coefficients were computed to assess the stability of frequency counts across the 1,000-word samples. Table 2 presents Cronbach's Alpha coefficients for the text sub-samples taken from spoken texts (the London-Lund corpus). In this case, four 1,000-word samples are analyzed from each spoken text. The alpha coefficient for each feature represents the average correlation among the four frequency counts of that feature (i.e. a count for each of the sub-samples); the standardized item alpha represents the alpha value that would be obtained if all of the frequency counts were standardized to have a variance of 1. Table 3 presents the reliability coefficients for the two 1,000-word sub-samples taken from each written text (the LOB corpus).

Table 1 Descriptive statistics for frequency scores (per 1,000 words) of ten linguistic features in a corpus of 481 texts taken from twenty-three spoken and written text genres.

Linguistic feature	Mean	Min.	Max.	Range
Nouns	181	84	298	214
Prepositions	111	50	209	159
Present tense	78	12	182	170
Past tense	40	0	119	119
Third person pronouns	30	0	124	124
First person pronouns	27	0	122	122
Contractions	14	0	89	89
Passives	10	0	44	44
WH relative clauses	3.5	0	26	26
Conditional subordination	2.5	0	13	13

Table 2 Reliability coefficients for four 1,000-word samples taken from spoken texts (conversations, broadcasts, and speeches); $N=9$ (i.e. 36 1,000-word samples; 4 samples from 9 texts)

First persons pronouns
ALPHA = 0.8878 STANDARDIZED ITEM ALPHA = 0.9226

Third person pronouns
ALPHA = 0.8946 STANDARDIZED ITEM ALPHA = 0.9029

Contractions
ALPHA = 0.9636 STANDARDIZED ITEM ALPHA = 0.9678

Present tense
ALPHA = 0.9431 STANDARDIZED ITEM ALPHA = 0.9483

Past tense
ALPHA = 0.8835 STANDARDIZED ITEM ALPHA = 0.8984

Nouns
ALPHA = 0.9594 STANDARDIZED ITEM ALPHA = 0.9652

Prepositions
ALPHA = 0.9359 STANDARDIZED ITEM ALPHA = 0.9371

Relative clauses
ALPHA = 0.9084 STANDARDIZED ITEM ALPHA = 0.9508

Passives
ALPHA = 0.7399 STANDARDIZED ITEM ALPHA = 0.7385

Conditional subordination
ALPHA = 0.7922 STANDARDIZED ITEM ALPHA = 0.7889

Table 3 Reliability coefficients for two 1,000-word samples taken from written texts (official documents, academic prose, general fiction, and romance fiction); $N=37$ (i.e. 74 1,000-word sub-samples; 2 samples from 37 texts)

First persons pronouns
ALPHA = 0.8277 STANDARDIZED ITEM ALPHA = 0.8290

Third person pronouns
ALPHA = 0.8593 STANDARDIZED ITEM ALPHA = 0.8641

Contractions
ALPHA = 0.8047 STANDARDIZED ITEM ALPHA = 0.8129

Present tense
ALPHA = 0.6108 STANDARDIZED ITEM ALPHA = 0.6123

Past tense
ALPHA = 0.9462 STANDARDIZED ITEM ALPHA = 0.9462

Nouns
ALPHA = 0.8963 STANDARDIZED ITEM ALPHA = 0.8964

Prepositions
ALPHA = 0.9503 STANDARDIZED ITEM ALPHA = 0.9503

Relative clauses
ALPHA = 0.5778 STANDARDIZED ITEM ALPHA = 0.5778

Passives
ALPHA = 0.9058 STANDARDIZED ITEM ALPHA = 0.9198

Conditional subordination
ALPHA = 0.3068 STANDARDIZED ITEM ALPHA = 0.3116

The alpha coefficients for the frequency counts of these features show a high level of reliability across the 1,000-word text samples (alpha higher than 0.85 for most features). The coefficients are generally smaller for the written samples (Table 3) than for the spoken samples, because they are based on two instead of four sub-samples. Only conditional subordination in the written texts has a low reliability coefficient (about 0.3), while present tense and relative clauses in the written texts have moderate reliability coefficients (about 0.6). With these exceptions, the reliability coefficients indicate that theses frequency counts are stable across the 1,000 word samples.[1]

Another way to assess the representativeness of the 1,000-word samples is to compute difference scores for pairs of samples from each text. Table 4 presents these difference scores for the same ten linguistic features and seven text categories as above. (Part A presents difference scores for the pronouns, contractions, and tense features; Part B presents the scores for nouns, prepositions, relative clauses, passives, and conditional subordinators). The scores in this table represent difference scores as a percentage of the total range of variation for the feature in question (given in Table 1). For example, the total range of variation for first person pronouns is 122 (in this case because some texts had 0 pronouns while others had 122—see Table 1). In conversation, the mean difference score for first person pronouns was nineteen, which thus represents a 16% difference as shown in Table 4 (i.e. 19/122 = 0.16). Presenting the difference scores as percentages permits direct comparisons across features, whereas differences in raw frequency do not. For example, a difference of ten first person pronouns is only 8% of their total range of variation (10/122 = 0.08), while a difference of ten WH relative clauses is a much higher 38% of their total range of twenty-six (i.e. 10/26 = 0.38). Table 4 presents the mean percentage difference score for each feature in each text category, plus the minimum and maximum difference scores for the text category. For example, the sub-samples from conversation texts show on average a 16% difference in first person pronouns, with a minimum difference of 10% and a maximum difference of 27%.

Table 4 shows that most of the difference scores are quite small across sub-samples. Over half of the difference scores are under 10%, and nearly all of the difference scores are less than 15%. Thus the overall characterization from Table 4 confirms the stability of the feature counts across 1,000-word sub-samples. It is interesting, though, to focus on the feature counts that show relatively large differences, especially to the extent that there are systematic patterns. Among the text categories, conversation and general fiction show rela-

Table 4 Descriptive statistics for difference scores of linguistic features between 1,000-word sub-samples taken from spoken and written texts; presented as a percentage of the total range of variation for the feature in question. $N = 55$ pairs (i.e. 110 1,000-word sub-samples)

PART A: PRONOUNS, CONTRACTIONS, AND TENSE FEATURES

GENRE	First person pronouns	Third person pronouns	Contractions	Past tense	Present tense
Conversations (N=8)					
Mean	16	11	9	19	11
Min.	10	0	2	9	1
Max.	27	31	25	39	23
Broadcasts (N=4)					
Mean	5	5	8	8	5
Min.	2	1	1	0	2
Max.	7	12	16	22	9
Speeches (N=6)					
Mean	14	7	8	9	12
Min.	0	2	1	0	1
Max.	38	17	12	27	22
Official documents (N=7)					
Mean	2	3	0	4	6
Min.	0	0	0	0	1
Max.	16	5	0	11	12
Academic prose (N=10)					
Mean	2	4	0	11	9
Min.	0	0	0	1	2
Max.	9	15	0	38	21
General fiction (N=10)					
Mean	12	23	15	15	14
Min.	3	0	0	2	0
Max.	32	55	43	34	40
Romance fiction (N=10)					
Mean	11	18	7	9	8
Min.	0	2	0	2	1
Max.	41	57	26	18	15

PART B: SELECTED NOMINAL AND STRUCTURAL FEATURES

GENRE	Nouns	Prepositions	Relative clauses	Passives	Conditional subordination
Conversations					
Mean	3	8	6	7	17
Min.	0	1	0	2	0
Max.	11	25	21	14	46
Broadcasts					
Mean	11	5	3	7	8
Min.	1	1	0	5	0
Max.	19	9	7	9	15
Speeches					
Mean	6	11	10	12	8
Min.	0	1	3	7	0
Max.	13	24	28	20	15
Official documents					
Mean	4	11	21	11	10
Min.	0	4	10	2	0
Max.	8	19	34	34	46
Academic prose					
Mean	9	5	14	21	11
Min.	1	1	0	0	0
Max.	33	18	31	52	23
General fiction					
Mean	6	8	5	6	24
Min.	0	0	0	0	0
Max.	19	21	14	27	77
Romance fiction					
Mean	6	5	5	6	19
Min.	0	0	0	2	8
Max.	15	16	24	11	46

tively large differences with respect to the pronominal, contracted, and tense features. In both of these cases, these differences probably reflect changing purposes within the course of a text; for example, shifts from monologic narrative to interactive discussion within conversation, or shifts from narrative to description to dialogue within fiction. In the case of fiction, it is interesting to compare general fiction to romance fiction, where the former shows consistently larger difference scores than the latter (possibly due to greater stylistic variety in general fiction). Conditional subordination also has high mean difference scores in conversation, general fiction and romance fiction, reflecting the low reliability coefficient for this feature. The lack of stability of this feature is probably related to its relative rarity in English texts, since the occurrence of even two conditionals in a sub-sample would result in a 15% difference score. The distributions of relative clauses and passives are also influenced by this relative rarity, although these two features are considerably more stable than conditional subordinators. This probably reflects the different discourse functions of these features: relative clauses and passives are widely distributed characteristics of style relating to informational elaboration and packaging; conditional subordinators, on the other hand, can serve a much more localized function within the rhetorical organization or argument of a text.[2]

Overall, the results presented in Tables 2–4 indicate a high level of stability for these linguistic feature counts across 1,000-word sub-samples of texts. This stability holds generally across linguistic features and across text categories. Given this stability between 1,000-word samples, it seems safe to conclude that the 2,000-word and 5,000-word texts in the standard corpora are reliable representatives of their respective text categories for analyses of this type.

3. Issues regarding text categories

The text categories in the Brown, LOB, and London-Lund corpora represent folk 'genres' readily distinguished by mature speakers of English, such as novels, newspaper articles, editorials, academic articles, public speeches, radio broadcasts, and everyday conversations. These categories are defined primarily on the basis of format, purpose, and situational context. Thus, newspaper articles are found in the news sections of newspapers; academic articles are found in academic journals; academic lectures differ from academic articles in terms of the mode, setting, and possibility of interaction. In a series of studies (Biber, 1986, 1988; Biber and Finegan, 1989*a*), I have analyzed the linguistic characteristics of spoken and written genres in English; statistical tests in those studies show that there are significant and important linguistic differences among genres (Biber, 1988, 127).

Genre distinctions do not by themselves, however, fully represent the underlying text distinctions of English, since texts within a genre can differ markedly in their linguistic characteristics. For example, newspaper articles can range from narrative and colloquial in linguistic form to informational and elaborated in form. On the other hand, some genres can be quite similar linguistically; for example, newspaper articles and popular magazine articles can be very similar in linguistic form. Biber (1989) proposes using two complementary perspectives to analyze the text categories of English: one is a typology of genres, as above, and the other is a typology of 'text types', representing the linguistically well-defined text categories of English. Under this latter perspective, linguistically distinct texts within a genre would represent different text types, while linguistically similar texts from different genres represent a single text type.

Biber (1989) uses multivariate statistical techniques, including cluster analysis, to identify the salient linguistically defined text types of English (cf. Biber and Finegan, 1989*b*). Overall, eight text types are identified; each text type represents a grouping of texts that are markedly similar to one another with respect to their linguistic characterizations. The types are interpreted by considering their predominant linguistic features, the general communicative characteristics of the texts grouped in each type, and microanalyses of particular texts; and functional labels are proposed for each type, such as 'Informational Interaction', 'Learned Exposition', and 'Involved Persuasion'. Although these text types are interpreted functionally, they are identified strictly on linguistic grounds; genres, on the other hand, are identified using situational criteria but can be characterized linguistically.

One of the main issues regarding text categories in corpus design is that some researchers do not accept genre categories as 'well-defined' (e.g. 'studies in linguistic variation on the basis of the main corpora . . . have failed to recognize that "genre" is not a well-defined

concept', Oostdijk, 1988, 18). In part, this criticism seems to be based on the existence of sub-genres within some genre categories; for example, the Brown and LOB corpora include a number of sub-genre categories (such as political, sports, society, financial, cultural, and spot news within press reportage; and numerous academic disciplines within academic prose). In addition, the above criticism seems to reflect scepticism concerning the linguistic coherence of genre categories—that is, the overriding question is whether the genre categories are linguistically well-defined even though they are not determined on linguistic grounds.

This question can be addressed empirically. As noted above, genres can be characterized linguistically, and they are statistically different from one another; but genres also have relatively large ranges of linguistic variation internally (much larger than within text types), because they are not defined on linguistic grounds. For this reason, the linguistic characterization of a genre should include both its central tendency and its range of variation. In fact, some genres are similar in their central tendencies but differ markedly in their ranges of variation (e.g. science fiction versus general fiction, and official documents versus academic prose, where the first genre of the pair has a much more restricted range of variation). In Biber (1988, 170–98), I describe the linguistic variation within genres, including the linguistic relations among various sub-genres. In the following discussion I present additional data to address two specific questions: the extent to which linguistic counts are stable across texts within a genre, and the number of texts needed to reliably represent a genre.

These questions are addressed by comparing the mean frequency counts for sub-samples of texts from within particular genre categories (an approach similar to that used in Section 2 above). Differences among 10-text samples are compared with differences among 5-text samples to assess the relative reliabilities of each. The results indicate a high degree of stability for mean genre scores across both 10-text and 5-text samples, although there is a notable decrease in reliability across the latter.

Five genres in the LOB and London-Lund corpora contained a sufficient number of texts for differences among 10-text samples to be tested: conversations, public speeches, press reportage, academic prose, and general fiction. Three 10-text samples were extracted from each of these genres,[3] and the mean frequency counts of six linguistic features were compared across the samples. Table 5, which presents the reliability coefficients for these frequency counts across the three 10-text samples, shows an extremely high degree of stability for all six linguistic features (alpha greater than 0.95). These coefficients show that the mean scores of the 10-text samples are very highly correlated. Two inferences can be drawn from these figures: (1) since the central linguistic tendencies of genres are quite stable, these categories are linguistically well-defined, even though they are not defined on linguistic grounds and show large ranges of internal variation; and (2) 10-text samples are large enough to reliably represent a genre.

This same procedure was repeated with 5-text samples from the same genres. As shown in Table 6, the reliabilities of mean scores across these 5-text samples is decreased slightly from the reliabilities across 10-text samples, but all six coefficients still indicate a high degree of stability across samples (alpha greater than 0.90). Thus, even 5-text samples seem to enable good assessments of the central tendency of a genre.

Table 5 Reliability coefficients among mean frequency counts of three 10-text samples taken from conversations, speeches, press reportage, academic prose (see note 3), and general fiction. $N = 6$ (3 samples from 6 genres with 10 texts/sample = 180 texts)

First person pronouns	
ALPHA = 0.9747	STANDARDIZED ITEM ALPHA = 0.9781
Third person pronouns	
ALPHA = 0.9935	STANDARDIZED ITEM ALPHA = 0.9972
Past tense	
ALPHA = 0.9878	STANDARDIZED ITEM ALPHA = 0.9915
Nouns	
ALPHA = 0.9949	STANDARDIZED ITEM ALPHA = 0.9952
Prepositions	
ALPHA = 0.9932	STANDARDIZED ITEM ALPHA = 0.9944
Passives	
ALPHA = 0.9813	STANDARDIZED ITEM ALPHA = 0.9844

Table 6 Reliability coefficients among mean frequency counts of three 5-text samples taken from conversations, speeches, press reportage, academic prose, and general fiction. $N = 10$ (i.e. 2 sets of text samples from 5 genres–2 sets of 3 samples from 5 genres with 5 texts/sample = 150 texts)

First person pronouns	
ALPHA = 0.9541	STANDARDIZED ITEM ALPHA = 0.9586
Third person pronouns	
ALPHA = 0.9611	STANDARDIZED ITEM ALPHA = 0.9685
Past tense	
ALPHA = 0.9759	STANDARDIZED ITEM ALPHA = 0.9775
Nouns	
ALPHA = 0.9744	STANDARDIZED ITEM ALPHA = 0.9759
Prepositions	
ALPHA = 0.9638	STANDARDIZED ITEM ALPHA = 0.9706
Passives	
ALPHA = 0.8984	STANDARDIZED ITEM ALPHA = 0.9029

As noted above, however, the linguistic characterization of a genre depends on both its central tendency *and* its range of variation. Table 7 provides descriptive statistics of the means and standard deviations for the linguistic frequency counts in each 10-text and 5-text sub-sample considered above. For example, the first set of scores on Table 7, Part A, gives descriptive statistics for the 5-text subsamples taken from the conversation genre. The first column describes the mean scores for first person pronouns: the minimum mean score in any sub-sample is 50; the maximum mean score is 65; the mean of the mean scores from all sub-samples is 59; and the standard deviation among the mean scores is 5. The second column describes the standard deviations for first person pronouns: the minimum standard deviation in any sub-sample is 5; the maximum standard deviation is 20; the mean of the standard deviations from all sub-samples is 12; and the standard deviation of the standard deviation scores is 6.

Table 7 is organized by genre, with the statistics for the 5-text samples presented before those for the 10-text samples. The table also includes statistics for 5-text samples from official documents and romance fiction (for comparison with academic prose and general fiction respectively). Part A presents the statistics for first person pronouns, third person pronouns, and past tense verbs, and Part B presents the statistics for nouns, prepositions, and passive forms.

The mean scores in Table 7 (i.e. mean of the mean scores and mean of the standard deviations) simply record the overall mean and standard deviation for the genre in question. Since the 5-text and 10-text samples are taken from the same genres, any differences between them is due to rounding. The more interesting statistics are the range indicators (Std, Min, Max, and Range); these show how much the mean scores and standard deviations of each sub-sample differed from the overall average mean and standard deviation. A quick look through Table 7 shows that the 5-text samples consistently have a wider range of deviation than the 10-text samples; and that the 10-text samples consistently show quite small ranges of deviation. The 10-text samples thus consistently provide good representations of these genres with respect to both their overall central tendency (the mean score—see Table 5) and their overall range of variation (the standard deviation—Table 7). The 5-text samples provide a fairly stable representation of some genres (e.g. academic prose and romance fiction for Part A features, and conversation and romance fiction for Part B features), but relatively unstable representations of other genres (e.g. public speeches for Part A features; academic prose for Part B features). In sum, the results presented in Tables 5 and 7 show that 10-text (and to a lesser extent 5-text) sub-samples accurately represent the linguistic characteristics of genre categories, including both the central tendency and the range of variation; this generalization holds even when the genre category contains a wide range of variation. These findings indicate that researchers are justified in using genres as a basis for variation studies, and that the coverage of most categories in the standard corpora, which typically include anywhere between twenty and eighty texts per category, is adequate for these types of analyses.

4. Issues regarding the overall size and composition of a corpus

Two other major issues relating to corpus design are: how big a corpus should be, and how diverse the texts in a corpus should be. For lexicographic research, several researchers have argued that corpora much larger than the LOB, Brown and London-Lund are required to identify occurrences of rare words, rare senses of common words, and rare collocational patterns (see, for example, Renouf, 1987). The same argument can be made for analyses of rare syntactic constructions (e.g. certain types of cleft constructions). With respect to variation studies, Oostdijk (1988) states that we do not know at present how large a corpus must be to provide the basis for meaningful statistical analyses (p. 20); but that new corpora should be less diverse than the Brown, LOB, and London-Lund, because corpora that represent 'a cross-section of a language . . .neutralize any variety-specificity' and thus 'are not suited for the study of linguistic variation' (p. 17). In the present section, I argue that the opposite priorities are better suited for variation studies; that is, corpora the size of the existing text corpora are adequate for correlation-based statistical analyses of linguistic variation, *if* those corpora represent the full range of text variation.

In Biber (1988) I analyze the overall parameters of variation among English texts. The study is based on a corpus of 481 texts (totalling approximately one million words), representing all of the major text categories in the LOB and London-Lund corpora plus collections of personal and professional letters. Using a factor analysis of the frequency counts of sixty-seven linguistic features in each text, five major parameters of variation are identified. Each parameter represents a distinct grouping of linguistic features that co-occur frequently in texts. On the assumption that linguistic co-occurrence reflects shared functions, each parameter is interpreted functionally (suggesting labels such as 'Involved versus Informational Production', and 'Elaborated versus Situated Reference').

Two questions regarding this study are relevant to the present discussion: should the study have been based on a much larger corpus of texts, and could/should the study have been based on a more restricted range of textual variation, analyzing fewer text categories but more texts per category? At present, it is not possible to compare the results of the above study with analyses of larger corpora, because such corpora (tagged for grammatical category) are not available. It is possible, however, to analyze various subcorpora extracted from the corpus used in the above study, and to compare those analyses with the analysis of the full corpus. Both the size and the representativeness of these subcorpora can be varied. If smaller subcorpora reliably represent the parameters of variation found in the analysis of the full corpus, then it is safe to assume that the original corpus is large enough for this purpose. Similarly, if subcorpora with a restricted range of variation reliably represent the parameters of variation, then the desirability of such corpora will be supported. (Without the availability of broader corpora, it is not possible to investigate the additional question of whether a *fuller* range of variation is desirable.)

Seven sub-corpora were extracted from the original corpus: six representing the full range of variation and one representing a restricted range of variation. First, the corpus was split in half with alternate texts being put into groups. Thus each of the two sub-corpora had 240 texts but represented the full range of genres contained in the original corpus. This same procedure was repeated a second time dividing the original corpus into four parts; in this case each sub-corpus had only 120 texts but again represented all genre categories. Factor analyses were run on each of these six sub-corpora and compared

Table 7 Descriptive statistics for the means and standard deviations of linguistic feature counts in 5-text and 10-text sub-samples from conversations, speeches, press reportage, academic prose, official documents, general fiction, and romance fiction.

PART A: 1ST PERSON PRONOUNS, 3RD PERSON PRONOUNS, PAST TENSE

Genre	Mean— 1st person pronouns	Std dev.— 1st person pronouns	Mean— 3rd person pronouns	Std dev.— 3rd person pronouns	Mean— past tense	Std dev.— past tense
Conversations—5-text samples (N=6; 30 texts)						
Mean	59	12	31	16	39	18
Std dev.	5	6	4	6	8	5
Min.	50	5	24	10	25	9
Max.	65	20	38	23	46	23
Range	15	15	14	13	21	14
Conversations—10-text samples (N=3; 30 texts)						
Mean	57	13	31	17	36	17
Std dev.	6	2	3	4	2	1
Min.	51	12	27	13	33	16
Max.	63	15	33	21	38	18
Range	12	3	6	8	5	2
Speeches—5-text samples (N=6; 30 texts)						
Mean	53	27	33	16	57	34
Std dev.	12	15	7	7	11	7
Min.	42	13	21	2	41	27
Max.	73	54	42	23	70	44
Range	31	41	21	21	29	17
Speeches—10-text samples (N=3; 30 texts)						
Mean	53	25	34	18	57	34
Std dev.	13	14	0	1	8	3
Min.	44	16	34	17	53	30
Max.	67	41	35	20	66	36
Range	23	25	1	3	13	6
Press reportage—5-text samples (N=6; 30 texts)						
Mean	9	6	30	10	46	18
Std dev.	5	4	8	6	5	8
Min.	3	2	15	4	37	10
Max.	17	11	36	20	53	29
Range	14	9	21	16	16	19
Press reportage—10-text samples (N=3; 30 texts)						
Mean	9	8	28	13	45	20
Std dev.	1	1	0	0	3	3
Min.	8	7	28	13	41	16
Max.	11	9	28	13	47	22
Range	3	2	0	0	6	6
Academic prose—5-text samples (N=6; 30 texts)						
Mean	9	9	10	8	15	17
Std dev.	3	3	3	3	5	10
Min.	5	4	6	4	10	8
Max.	14	13	16	12	24	35
Range	9	9	10	8	14	27
Academic prose—10-text samples (N=6; 60 texts)						
Mean	6	7	11	10	22	22
Std dev.	1	2	2	2	4	5
Min.	4	4	9	6	18	18
Max.	8	9	14	13	30	31
Range	4	5	5	7	12	13
Official documents—5-text samples (N=6; 30 texts)						
Mean	11	12	10	5	16	13
Std dev.	9	10	1	1	8	9
Min.	1	1	9	4	8	5
Max.	17	20	11	6	23	22
Range	16	19	2	2	15	17

Table 7 PART A *(continued)*

Genre	Mean— 1st person pronouns	Std dev.— 1st person pronouns	Mean— 3rd person pronouns	Std dev.— 3rd person pronouns	Mean— past tense	Std dev.— past tense
General fiction—5-text samples (N=6; 30 texts)						
Mean	32	23	68	28	85	15
Std dev.	9	8	6	9	8	7
Min.	20	13	58	16	72	10
Max.	43	32	77	37	93	30
Range	23	19	19	21	21	20
General fiction—10-text samples(N=3; 30 texts)						
Mean	32	24	67	27	86	15
Std dev.	3	2	5	5	5	5
Min.	29	22	61	22	80	11
Max.	35	26	70	32	90	21
Range	6	4	11	10	10	10
Romance fiction—5-text samples (N=6; 30 texts)						
Mean	32	10	79	13	84	11
Std dev.	7	2	4	8	5	3
Min.	26	8	75	4	81	8
Max.	40	12	84	19	89	14
Range	14	4	9	15	8	6

PART B: NOUNS, PREPOSITIONS, PASSIVES

Genre	Mean— nouns	Std dev.— nouns	Mean— prepositions	Std dev.— prepositions	Mean— passives	Std dev.— passives
Conversations—5-text samples (N=6; 30 texts)						
Mean	136	15	84	11	4	2
Std dev.	5	3	3	2	1	1
Min.	130	12	80	9	3	1
Max.	144	19	89	14	6	4
Range	14	7	9	5	3	3
Conversations—10-text samples (N=3; 30 texts)						
Mean	137	16	85	11	4	2
Std dev.	2	1	4	1	1	1
Min.	136	15	82	10	4	2
Max.	140	17	90	12	5	3
Range	4	2	8	2	1	1
Speeches—5-text samples (N=6; 30 texts)						
Mean	173	26	104	20	8	4
Std dev.	5	9	7	5	2	2
Min.	166	11	94	14	5	2
Max.	179	35	112	26	10	6
Range	13	24	18	12	5	4
Speeches—10-text samples (N=3; 30 texts)						
Mean	172	26	103	20	8	4
Std dev.	3	3	4	5	1	1
Min.	170	22	101	15	6	4
Max.	175	28	107	24	9	5
Range	5	6	6	9	3	1
Press reportage—5-text samples (N=6; 30 texts)						
Mean	222	18	116	13	12	5
Std dev.	8	6	7	4	4	2
Min.	210	10	111	8	8	2
Max.	234	24	127	17	20	7
Range	24	14	16	9	12	5
Press reportage—10-text samples (N=3; 30 texts)						
Mean	221	18	116	13	13	5
Std dev.	6	10	2	4	1	1
Min.	217	12	115	8	12	4
Max.	228	30	118	16	14	6
Range	11	18	3	8	2	2

Table 7 PART B *(continued)*

Genre	Mean—nouns	Std dev.—nouns	Mean—prepositions	Std dev.—prepositions	Mean—passives	Std dev.—passives
Academic prose—5-text samples (N=6; 30 texts)						
Mean	181	26	143	16	18	7
Std dev.	14	10	9	9	3	5
Min.	157	13	132	4	14	1
Max.	193	42	158	31	22	15
Range	36	29	26	27	8	14
Academic prose—10-text samples (N=6; 60 texts)						
Mean	190	23	139	17	19	8
Std dev.	4	6	4	5	2	2
Min.	184	16	135	12	17	6
Max.	195	33	146	27	23	10
Range	11	17	11	15	6	4
Official documents—5-text samples (N=6; 30 texts)						
Mean	206	15	151	23	21	7
Std dev.	20	9	15	5	3	4
Min.	189	7	138	18	19	3
Max.	228	24	167	26	23	11
Range	39	17	29	8	4	8
General fiction—5-text samples (N=6; 30 texts)						
Mean	161	27	93	16	6	3
Std dev.	7	8	5	7	1	1
Min.	148	20	86	7	5	2
Max.	168	38	102	25	8	4
Range	20	18	16	18	3	2
General fiction—10-text samples(N=3; 30 texts)						
Mean	161	26	93	16	6	3
Std dev.	6	4	1	3	0	0
Min.	156	21	92	13	5	3
Max.	167	30	94	19	6	4
Range	11	9	2	6	1	1
Romance fiction—5-text samples (N=6; 30 texts)						
Mean	147	17	82	9	5	1
Std dev.	7	5	3	2	0	0
Min.	139	12	79	6	5	1
Max.	151	23	84	11	5	2
Range	12	11	5	5	0	1

with the dimensions identified on the basis of the full corpus. Finally, a relatively large sub-corpus with a restricted range of variation was extracted from the original corpus. This sub-corpus represented only expository, written texts (i.e. only the genres of academic prose, official documents, press editorials, press reviews, religion, and skills and hobbies). With 169 texts, this sub-corpus was larger than the quarter sub-corpora, and with all texts being expository in nature, it had greater depth and less diversity than any other sub-corpus.

Table 8 summarizes the factor analyses of each of these sub-corpora. The table is based on the five major factors identified from the original corpus; the features that 'load' on these five factors are identified by a '*' in the table. If a feature loads on a factor, it means that it shares an important amount of co-variation with the total co-occurrence relations represented by the factor. All features with loadings over 0.30 are listed in the table. Factors have negative and positive loadings, representing two groups of features that occur in a largely complementary distribution; the positive loadings are listed above the dashed line for each factor, and the negative loadings are listed below the dashed line.

In Table 8, the features loading on the factor analysis (FA) for each sub-corpus are marked by a numeral on the right side of the table. '1' marks the FA for the first half of the split corpus (240 texts); '2' for the second half of the split corpus; '3' marks the FA for one of the quarter corpora (120 texts); and '4' marks the FA for the corpus of 169 texts having a restricted range of variation. The factors produced by each of these analyses should be compared with the '*' features representing the original FA (based on all 481 texts).

It can be seen from Table 8 that the FAs of the two half corpora are good replications of the factorial structure of the original FA. Factors 1 and 2 are almost exact replications, while Factors 4 and 5 very closely replicate the factorial structure of the original study. (The FA of the first half corpus adds two features for both Factor 4 and Factor 5, but otherwise the structure of these two factors is also nearly identical to the original structure.) Only Factor 3 shows any notable difference, in that the

Table 8 Summary of the factorial structure of five factor analyses. *=whole corpus (481 texts); 1, 2=two halves of corpus (240 texts each); 3=one-fourth of corpus (120 texts); 4=169 texts restricted to informational prose. (All loadings over 0.3 are included.)

FACTOR 1					
Private verbs	*	1	2	3	
THAT deletion	*	1	2	3	4
Contractions	*	1	2	3	4
Present tense verbs	*	1	2	3	4
2nd person pronouns	*	1	2	3	4
DO as pro-verb	*	1	2	3	4
Analytic negation	*	1	2	3	
Demonstrative pronouns	*	1	2	3	
General emphatics	*	1	2	3	
First person pronouns	*	1	2	3	4
Pronoun IT	*	1	2	3	
BE as main verb	*	1	2	3	
Causative subordination	*	1	2	3	
Discourse particles	*	1	2	3	
Indefinite pronouns	*	1	2	3	4
General hedges	*	1	2	3	
Amplifiers	*	1	2	3	
Sentence relatives	*	1	2	3	
WH questions	*	1	2	3	
Possibility modals	*	1	2	3	
Non-phrasal coordination	*	1	2	3	4
WH clauses	*	1	2		
Final prepositions	*	1	2	3	4
Adverbs	*	1	2	3	
Conditional subordination	*	1	2		
Predicative adjectives				3	
3rd person pronouns					4
Nouns	*	1	2	3	
Word length	*	1	2	3	4
Prepositions	*	1	2	3	4
Type/token ratio	*	1	2	3	
Attributive adjectives (no*)	*	1	2	3	4
Place adverbials	*	1			
Agentless passives	*	1	2	3	
Past participial WHIZ deletions	*	1	2		
Present participial WHIZ deletions	*	1	2		
BY-passives		1			
Phrasal coordination			2	3	
Nominalizations					4
FACTOR 2					
Past tense verbs	*	1	2	3	
Third person pronouns	*	1	2	3	4
Perfect aspect verbs	*	1	2	3	4
Public verbs	*	1	2	3	
Synthetic negation	*	1	2	3	
Present participial clauses	*		2		
General emphatics					4
BE as main verb					4
Type/token ratio					4
Pronoun IT					4
Attributive adjectives					4
Time adverbials					4
Adverbs					4
Present tense verbs	*	1	2	3	
Attributive adjectives	*	1	2	3	
Prepositions				3	4
Nominalizations				3	4
Past participial WHIZ deletions				3	4
Agentless passives				3	4
BY-passives				3	4
Present participial WHIZ deletions					4
Past participial clauses					4

Table 8 *(continued)*

FACTOR 3					
WH relative clauses on object positions	*	1	2	3	4
Pied piping constructions	*	1	2	3	4
WH relative clauses on subject positions	*	1	2	3	4
Phrasal coordination	*				
Nominalizations	*		2		
THAT relative clauses on object positions		1	2		
Existential THERE		1			
THAT complement clauses			2		
— — — — — — — — — — — —					
Time adverbials	*				
Place adverbials	*				
Adverbs	*				
FACTOR 4					
Infinitives	*	1	2	3	4
Prediction modals	*	1	2	3	4
Suasive verbs	*	1	2		4
Conditional subordination	*	1	2		4
Necessity modals	*	1	2		4
Split auxiliaries	*	1		3	4
Possibility modals	*				4
THAT relative clauses on object positions		1		3	
THAT complement clauses		1		3	
Present tense					4
Predicative adjectives					4
— — — — — — — — — — — —					
Phrasal coordination				3	
Past tense					4
FACTOR 5					
Conjuncts	*	1	2	3	
Agentless passives	*	1	2	3	
Past participial clauses	*	1	2	3	
BY-passives	*	1	2	3	
Past participial WHIZ deletions	*	1	2		
Other adverbial subordinators	*	1		3	
Nominalizations	*	1	2		
Present participial WHIZ deletions		1		3	
Predicative adjectives				3	
— — — — — — — — — — — —					
Time adverbials	*	1	2	3	
Place adverbials			2		
Adverbs		1		3	

negative loadings are not replicated by either sub-corpus. Overall, the original FA is strongly confirmed, and the FAs of the two half-corpora indicate that a corpus of 240 texts fairly well represents the underlying parameters of variation.

Even more surprising is the extent to which the FA of the quarter sub-corpus (120 texts) replicates the original factorial structure. Again there is an extremely high degree of replication for Factor 1. Factors 2 and 5 are also replicated to a high degree, except that some additional features are included in the factorial structure of these factors. Factors 3 and 4 are less well replicated, although the basic structure of each factor is still apparent. Five of the original features are missing on Factor 3, while on Factor 4, two original features are missing and three features are added. Overall, though, these findings indicate that even a corpus of 120 texts enables a rough identification of the basic parameters of variation.

In contrast, the corpus of 169 expository texts, although larger than the quarter sub-corpus, provides a much worse representation of the underlying parameters of variation. Only Factor 4 is correctly replicated by this FA (with the addition of three features), while Factor 3 is replicated equally well by this FA and by the FA of the quarter sub-corpus. With respect to the other three factors, however, the basic factorial structure is not well-represented. Well over half of the features loading on the original Factors 1 and 2 are not identified by this FA, while fourteen additional features are included on Factor 2. Factor 5 is not identified at all by this analysis.

The poor replication of this last FA, based on a restricted sub-corpus, simply reflects a basic fact of bivariate correlation—that meaningful correlations must be based on a full range of variation (Welkowitz *et al.*, 1976, 162). Factor analysis is, in a sense, a multi-variate form of correlation, so these same qualifications hold. A more transparent example might be the inappropriateness of a psychologist attempting to identify the basic parameters of human intelligence by analyzing only the performance on math tests of advanced calculus students. Similarly, any attempt to identify the basic parameters of linguistic variation among texts based on analysis of a restricted corpus is inappropriate. That is, one of the most important considerations in designing a text corpus for multivariate statistical analyses is to ensure that it *does* represent a cross-section of the language in question.

In sum, this section has shown that the underlying parameters of text-based linguistic variation (as represented by the factorial structure) can be replicated in a relatively small corpus, *if* that corpus represents the full range of variation. In contrast, larger corpora are not adequate for overall analyses of textual variation if they fail to represent the range of variation. These analyses indicate that the total number of texts included in the existing computer-based corpora are adequate for multivariate statistical analyses; additional research is required to determine the extent to which existing corpora actually represent the full range of variation among texts in English.

5. Summary and conclusion

This paper should not be construed as arguing that 'small is beautiful' and that the existing corpora are perfectly designed. Rather, my purpose has been to show that existing corpora are adequate in many respects; in particular, that relatively short text lengths and small corpus size are often adequate, that genres are well-defined text categories, and that the design goal of representing a wide range of variation (adopted by existing corpora) is necessary if a corpus is to be used for analyses of textual variation. A number of issues relating to corpus design have not been addressed in the present study, including: (1) the extent to which the selection of texts within a genre actually represents the range of variation in English for that genre; (2) the extent to which the selection of genres (and the full range of texts) actually represents the full range of variation among texts in English; and (3) the suitability of existing corpora for research purposes other than descriptive and sociolinguistic analyses of linguistic variation. Investigation of these additional issues will help guide the construction of larger and more representative corpora.

My main point in the present paper, though, is that descriptive linguists should not be intimidated by the 'need' for larger corpora. Many as yet unaddressed linguistic issues could be profitably investigated taking a corpus approach; the existing computer-based corpora provide an important resource for the study of linguistic variation, and that resource has not been fully exploited to date. I have argued in the present paper that the existing corpora are adequate for many analyses of: the distribution of linguistic features, the linguistic characteristics of genres and text types, and the macroscopic parameters of variation. Given the increased scope and sophistication of analysis enabled by a corpus approach, together with the availability and demonstrated suitability of existing corpora, there is every reason to make maximal use of these corpora for analyses of linguistic variation until larger corpora become readily available.

Acknowledgements

I would like to thank Dwight Atkinson and Edward Finegan for their helpful comments on earlier drafts of this paper.

Notes

1. The low reliability of conditional subordination in the written texts is due in part to the low frequency counts of conditionals generally.
2. A similar kind of skewing might also hold for other specialized features, such as concessive adverbial subordinators or cleft constuctions, as well as for specialized uses of more general features, such as the use of first person pronouns in scientific articles.
3. There were enough academic prose texts in the LOB Corpus to use two separate sets of three 10-text samples from this genre (i.e. six 10-text samples in all), resulting in $N=6$ on Table 5.

References

Altenberg, B. (1984). 'Causal Linking in Spoken and Written English', *Studia Linguistica*, **38**, 20–69.

Biber, D. (1986). 'Spoken and Written Textual Dimensions in English: Resolving the Contradictory Findings', *Language*, **62**, 384–414.

—— (1988). *Variation Across Speech and Writing*. Cambridge: Cambridge University Press.

—— (1989). 'A Typology of English Texts', *Linguistics*, **27**, 3–43.

—— and Finegan, E. (1989*a*). 'Drift and the Evolution of English Style: A History of Three Genres', *Language*, **65**, 487–517.

—— —— (1989*b*). 'Styles of Stance in English: Lexical and Grammatical Marking of Evidentiality and Affect', *Text*, **9**, 93–124. (Special issue on the pragmatics of affect, edited by Elinor Ochs.)

Coates, J. (1983). *The Semantics of the Modal Auxiliaries*. London: Croom Helm.

Francis, W. and Kucera, H. (1982). *Frequency Analysis of English Usage: Lexicon and Grammar*. Boston: Houghton Mifflin.

Garside, R., Leech, G., and Sampson, G. (eds.) (1987). *The Computational Analysis of English: A Corpus-based Approach*. London: Longman.

Johansson, S. (ed.) (1982). *Computer Corpora in English Language Research*. Bergen: Norwegian Computing Centre for the Humanities.

—— and Hofland, K. (1989). *Frequency Analysis of English Vocabulary and Grammar*. Oxford: Oxford University Press.

Kucera, H. and Francis, W. Nelson. (1967). *Computational Analysis of Present-day American English*. Providence, RI: Brown University Press.

Oostdijk, N. (1988). 'A Corpus Linguistic Approach to Linguistic Variation'. *Literary and Linguistic Computing*, **3**, 12–25.

Renouf, A. (1987). 'Lexical Resolution'. In W. Meijs (ed.), *Corpus Linguistics and Beyond*, 121–131. Amsterdam: Rodopi.

Sinclair, J. M. (ed.) (1987). *Looking Up: An Account of the COBUILD Project in Lexical Computing*. Glasgow: Collins.

Svartvik, J. and Quirk, R. (eds.) (1980). *A Corpus of English Conversation*. Lund Studies in English 56. Lund: Lund University Press.

Tottie, G. and Bäcklund, I. (eds.) (1986). *English in Speech and Writing: A Symposium*. Studia Anglistica Upsaliensia 60. Stockholm: Almqvist and Wiksell.

Walker, D. E. (1989). 'Developing Lexical Resources', *Proceedings of the 5th Annual Conference of the UW Centre for the New Oxford English Dictionary*, 1–22. Waterloo, Ontario: University of Waterloo Centre for the New Oxford English Dictionary.

Welkowitz, J., Ewen, R. E. and Cohen, J. (1976). *Introductory Statistics for the Behavioral Sciences*. New York: Academic Press.