CHAPTER

10

Corpus Linguistics and Language Variation

Michael P. Oakes

The creation of balanced corpora representing national varieties of English throughout the world has enabled statistical comparisons of word frequencies in each of these corpora. This chapter examines some of the statistical techniques that have been used in order to carry out such comparisons. Starting with the chi-squared test, I examine studies carried out on the Brown family of corpora, and note how use of dispersion measures (Juilland et al. 1970) and the Bonferroni correction (Altman 1991) can improve such comparisons. I then examine studies of style, genre and register variation, which have used multivariate techniques such as factor analysis and hierarchical clustering. Automatic genre differentiation by a Support Vector Machine (SVM) will be described, not with the objective of making new linguistic discoveries, but because automatic genre identification is useful in itself – in this case as a component in genre-sensitive search engines. Finally, the chapter concludes with a discussion of difficulties of isolating single factors in corpora for comparison.

10.1 Introduction

Hofland and Johansson (1982: 39) state that quantitative studies of linguistic variation 'shed new light on the utilization of words in different varieties of English, and can, in addition, serve as a starting point for stylistic and grammatical studies as well as for cultural observations'. Biber (1988: 3) writes that many linguists are interested in differences between speech and writing, which can be considered as poles of one dimension of linguistic variation. However, certain types of stylistic research are not amenable to computer analysis, either because they involve a good deal of expert intuition (such as the identification of Semitisms in the Greek New Testament), or because they consider linguistic features which are found only rarely, such as the hapax legomena (words which appear only once) in a text, although these words tell us much about the writer's background and experience. Computer analyses of linguistic variation thus tend to be restricted to comparisons of the use of frequently occurring, objectively countable linguistic features, which are focussed on in this chapter.

Variationist research questions often take the form of 'what are the differences in the ways that feature x is used in corpus y and corpus z?' It should be noted that such questions also implicitly address similarities (although researchers tend to

find differences more interesting to report), and that multiple (rather than two) corpora can be compared (as discussed later in this chapter). Additionally, the examination of differences is usually frequency-based (e.g. how often does feature x occur in corpus y and is this significantly more or less than in corpus z), although frequencies can often be approached from more complex perspectives, for example, if feature x has three specific functions a, b and c, how are these functions proportionally represented in corpora y and z?

Clearly, in order to carry out comparisons, it is important to use corpora that are matched in as many ways as possible to reduce the number of independent variables that could impact on variation. For example, if a researcher wanted to compare the spoken language of males with the language of females, then he/she would try to gather two corpora (male speech and female speech) that were of similar sizes, had similar numbers of speakers who were talking in similar contexts, in similar time periods and in similar regions of the world. Clearly, the more variables that are different between the two corpora being studied, the more difficult it becomes to say that any differences are the result of the factor we want to examine (such as sex). The Brown corpus family, discussed below, are a good example of corpora that have been carefully constructed using the same sampling model to carry out studies of diachronic and synchronic variation.

10.1.1 The Brown Corpus Family

Before examining the various statistical techniques which have been used to examine corpus variation, I will briefly review the well-known family of corpora based on the original Brown corpus, consisting of a million words of written American English texts that had been published in 1961 (Francis and Kučera 1964). The LOB (Lancaster-Oslo/Bergen) corpus is the British English equivalent of the Brown corpus, published in 1961. Hofland and Johansson (1982: 22) point out that 'One of the major advantages of the LOB and Brown corpora is that their composition enables a comparison of the characteristics of different types of texts'. The two corpora are 'balanced', in that the material in the two corpora were selected to match as much as possible. Both corpora consist of 500 text samples of about 2,000 words each, making about a million words in total. Table 10.1 shows the breakdown of text categories that are represented in both the LOB and the Brown corpora.

The corpora were assembled using identical sampling procedures. The whole population of texts were obtained from three published compendia, one consisting of samples from books, one taken from newspapers and periodicals and one consisting of government documents. Titles from these sources were chosen using a random number table, and articles obviously not written by British authors were excluded. Having chosen each title, the next task was to randomly select the page at which to start the 2,000 word extract. Non-random criteria were also used, such as excluding articles mainly consisting of dialogue, and weighting the sampling process so that articles from the national press were more likely to be chosen than articles from the local press (Hofland and Johansson 1982: 2).

Table 10.1 Text categories in the Brown family (data derived from Biber et al. 1998: 14)

	Broad text category	*Text category letter and description ('genre')*		*Number of texts*	
				Brown Frown	*LOB FLOB*
Informative	**Press**	**A**	Press: Reportage	44	44
		B	Press: Editorial	27	27
		C	Press: Reviews	17	17
	General Prose	**D**	Religion	17	17
		E	Skills, Trades and Hobbies	36	38
		F	Popular Lore	48	44
		G	Belles Lettres, Biographies, Essays	75	77
		H	Miscellaneous: Government documents, industrial reports etc.	30	30
	Learned Writing	**J**	Academic prose in various disciplines	80	80
Imaginative	**Fiction**	**K**	General Fiction	29	29
		L	Mystery and Detective Fiction	24	24
		M	Science Fiction	6	6
		N	Adventure and Western	29	29
		P	Romance and Love story	29	29
		R	Humour	9	9

10.1.2 Diachronic Corpora

Two other corpora, balanced with respect to LOB and Brown, are the FLOB (Freiberg-LOB) and Frown (Freiberg-Brown) corpora, designed to represent British English in 1991 and American English in 1992 respectively. Comparisons of these corpora are thus ideal for diachronic studies, enabling observation of changes in language over a thirty-year gap. More recently, further corpus building projects are underway: Leech and Smith (2005) have produced the Lancaster 1931 corpus of British English and are planning a 1901 version, while Baker (2008) has built the mid-2000s equivalent. Comparisons of these corpora will enable researchers to determine whether an observed linguistic change is speeding up, slowing down, remaining constant or reversing (although, it should be borne in mind that a gap of 30 or even 15 years may not give a full picture regarding linguistic change, but rather presents information as static 'snap-shots').

In contrast to the BNC Conversational Corpus (see Section 10.2.2), the Brown family corpora do not record any biographical or demographic information, since

this is often not known. However, the Brown corpus family have all been annotated with a common grammatical tagset known as C8 (Leech and Smith 2005).

10.1.3 Sources of Linguistic Variation

As shown in Table 10.1, the Brown family corpora are subdivided into categories, which correspond loosely to 'genre', 'text type' or 'register'. Biber (1998: 208) uses the term 'genre' for classes of texts that are 'determined on the basis of external criteria relating to author's or speaker's purpose'. On the other hand, he uses 'text type' to refer to classes of text that are grouped on the basis of 'similarities in intrinsic linguistic form, irrespective of their genre classifications'. For example, particular texts from press reportage, biographies and academic prose might be very similar in having a narrative linguistic form, and thus would be grouped together as a single text type, even though they represent different genres. Register (Biber 1998: 135) refers to variation in language arising from the situations it is used in, and depends on such things as its purpose, topic, setting, interactiveness and the intended addressee. Registers may be highly specific, such as novels by Jane Austen, or the 'Methods' sections of academic papers describing biological research. A fourth source of variation recognized by Biber is 'dialect', which is defined by association with different speaker groups, based on region, social group or other demographic factors. Genre can swamp changes in language change 'proper' as observed in diachronic studies. To quote Hundt and Mair (1999: 222) differences in balanced corpora may 'reflect a change in stylistic preference rather than a change in grammatical rules'.

10.1.4 Feature Selection

For many text classification tasks, an essential early stage is to decide which linguistic features (called 'attributes' in machine learning applications) should be used to characterize the texts. Often, texts are characterized by the single word tokens they contain. Another possibility is to reduce all the words in a text to their lemmas, and record how many of each lemma is found in that text. If the texts are annotated with a state-of-the-art semantic tagger (e.g. Rayson et al. 2004*a*), then they can be characterized by the frequency with which each semantic tag is found. Even when selecting single word tokens as attributes, some decisions must be taken. For example, Hofland and Johansson (1982: 7) specified that a word token should consist of 'alphanumeric characters surrounded by white space, but may contain punctuation', thus allowing 2.33 and 5,000 to be counted as words. They also devised a capitalization principle, incidentally providing a basis for the different task of named-entity or proper noun recognition: 'words which are spelled with capitals in all their occurrences in the material are reproduced with capitals in the lists'. Some words were given interpretive codes, for example, AB (abbreviation), SF (neologism from science fiction), FO (foreign words and expressions), so another basis for comparison between two varieties of English

could be the relative prevalence of these codes. For example, does British English make more use of abbreviations than American English?

One of the simplest and most widely used techniques for examining language variation in different corpora is the chi-squared test. The following section explains how the test is carried out and describes corpus variation studies which have used it. Additionally, I discuss limitations of the chi-squared test in order to demonstrate when its use is appropriate, and the advantage of using it in conjunction with a measure called Yule's Q.

10.2 The Chi-Squared Test

The chi-squared test is a widely used statistical test for the comparison of count data, such as word frequencies. One question which might be tackled by this test could be 'is the word mother more typical of female speech than male speech?' In the BNC Conversational Corpus, the word mother occurs 272 times in a corpus of 1,714,433 words of male speech, and 627 times in a 2,593,452 corpus of female speech (Rayson et al. 1977). The so-called 'observed' (directly counted) data is

(a) the word mother was spoken 627 times by females;
(b) mother was spoken 272 times by males;
(c) words other than mother were spoken 2,593,452 – 627 = 2,592,825 times by females, and;
(d) words other than mother were spoken on 1,714,433 – 272 = 1,714,161 occasions by males.

These values are set out in a 'contingency table', as shown in Table 10.2(a). However, do these numbers show that females really do say mother more than men? The first step in answering this question is to calculate the 'expected' values, which are the values for (a) to (d) we would have obtained if there were absolutely no difference in how often the two sexes say this word. The grand total, which is the total number of words in the combined corpora, is 1,714,433 + 2,593,452 = 4,307,885. Of this total, the word mother was spoken 272 + 627 = 899 times (the row total), and 2,593,452 words were spoken by men (the column total). So we might expect that of the 899 times that mother occurred, it would have been spoken by males 899 times the proportion of words in the combined corpus spoken by males, which is 899 × (1,714,433/4,307,885). Analogously, we work out the expected frequencies for the other three cells in the contingency table, using the formula

$$Expected\ \ frequency = \frac{Row\ total \times Column\ total}{Grand\ total}$$

The expected values are shown in Table 10.2(b). Now that we have corresponding tables of observed and expected values, we can derive a third table of contributions

Table 10.2 Calculation of chi-squared for the use of the word mother in male and female speech

(a) Observed values

	Male speech	*Female speech*	*Row total*
mother	272	627	899
any other word	1,714,161	2,592,825	4,306,986
Column total	1,714,433	2,593,452	*Grand Total* = 4,307,885

(b) Expected values

	Male speech	*Female speech*
mother	357.8	541.2
any other word	1,714,074.9	2,592,910.6

(c) Contributions to the overall chi-squared value made by each cell

	Male speech	*Female speech*
mother	20.6	13.6
any other word	0.0	0.0

to the overall chi-squared value made by each cell. The value in each cell of this table (shown in Table 10.2(c)) is given by the formula:

$$\frac{\left(\textit{Observed value for that cell} - \textit{Expected value for that cell}\right)^2}{\textit{Expected value for that cell}}$$

The four values in this table (one for each cell) are added together, to give an overall chi-square value (34.2 in this example). We then obtain a quantity called the degrees of freedom, which depends on the size of the original contingency table, as follows:

Degrees of freedom = (number of rows – 1) × (number of columns – 1)

Thus for a 2 by 2 table, we have one degree of freedom. Using this data, we can look up what are called the 'critical values' for the chi-squared statistic in tables which can be found at the back of many statistics textbooks, such as Table A5 in Woods et al. (1986). If chi-square is > 3.84 with one degree of freedom, we can be 95 per cent confident that there really is a difference in the number of times males and females say mother; if chi-squared > 6.64 we can be 99 per cent confident; and if chi-squared is > 10.8 (as is the case in the mother example), we can be 99.9 per cent confident.

The chi-squared test can be extended for comparisons of more than two corpora. For example, Hofland and Johansson (1982: 19) simultaneously compared word frequencies in LOB Brown, and two other corpora created by Carroll et al. (1971) and Jones and Sinclair (1974). Oakes and Farrow (2007) used the chi-squared test to compare the word frequencies in the variants of English used in Australia, Britain, India, New Zealand and the United States. Here the aim was to find words which occurred significantly more often in one variety of English than any of the others. The Australian corpus (ACE) had significantly more names of Australian people and places, and terms related to employment rights such as unions, unemployed and superannuation. As previously found by Leech and Fallon (1992) when comparing LOB and Brown, the British corpus (in this case FLOB) had many more aristocratic titles. The Kolhapur corpus of Indian English contained the terms crores for tens of millions, and lakhs for tens of thousands. There were also significantly more words related to the caste system, particularly caste, castes and dalit, and more terms coming from a wide variety of religions. The Wellington corpus of New Zealand English not surprisingly had many more names of New Zealand people and places, and also more sporting terms (rugby was sixteenth on the list of most typical New Zealand words) and words describing the natural world such as bay, beach and cliff. The Frown corpus had more terms reflecting concerns about diversity and equality: black, gender, white, diversity and gay. While many of these observations indicate cultural differences between these countries where English is widely spoken, linguistic differences were found as well. For example, the Kohlhapur corpus had more high frequency function words. Spelling differences between American English and the other types of English were regularly found, with color for example being highly typical of American English. Many concepts for transport were found to be typical of Frown, such as pickup, railroad, highway and transportation. These concepts exist in other English-speaking countries, but other terms are used for them.

When using the chi-squared test, a number of common errors should be avoided. First, some people feel that the test can only be used when comparing corpora of equal size. This is not so; it can be used for corpora of any relative size, as long as the rule-of-thumb is followed that expected values should all be five or more (for exact details of this, see Rayson et al. (2004*b*)). Second, some authors have disregarded data because some of the observed frequencies are less than five – but the criterion is that only words for which the expected values are less than five should be excluded from the analysis. Third, some authors have 'normalized' their data, so that instead of using raw frequency counts (the actual number of times each word was observed), they have expressed figures as ratios, for example, the occurrence of each word per thousand words of text. The problem with this is that different values of chi-squared would be obtained if the ratios were expressed as, say, words per million, so it is essential to use raw frequency counts rather than ratios, when carrying out chi-squared tests.

A potential problem regarding the value placed on results of the chi-squared tests relates to dispersion. For example, thalidomide occurs very frequently in the FLOB corpus, but only in a single article. Thus it would be dangerous to

conclude that this word is somehow very typical of the British English of the 1990s. Therefore, a number of authors have pointed out that it is important to take into account some sort of criterion of 'dispersion' when carrying out comparisons of word frequencies across corpora, so conclusions are not based on words which appear often, but are clumped into a very small number of texts. The criterion of Kučera and Francis (1967) for the Brown corpus is that words should be distributed through at least five texts. Finally, if we are making multiple comparisons, such as simultaneously comparing the frequencies of many different words in two corpora, there will be an increased risk of Type I errors. These occur when a statistical test shows that a significant difference has been found in the frequency of a word in two corpora, even though this difference was simply due to chance variation. If we are looking for significance at the 5 per cent level, a type I error will occur in about one in twenty of the comparisons we make. This problem can be compensated for by using the Bonferroni correction (Oakes and Farrow 2007). In the past, people (e.g. Dunning 1993) have argued that related measures such as log-likelihood, also known as G^2, can be used with very low frequency data. There are indications that the log-likelihood test is becoming more popular among corpus linguists (see, e.g. King, this volume). However, the 'minimum of 5 for Expected values' rule should still be used when using this test (Moore 2004).

Lucy (2005: 51) writes that 'one of the properties of the chi-squared test is that variables showing only a weak relationship can show highly significant values of chi-squared'. This is because the chi-squared value reflects how confident we can be that there really is a relationship between two variables (such as a word and a genre). If we repeat a chi-squared experiment with a much larger data set, we would be more confident of our findings, and the resulting chi-squared values would be greater than before. However, the strength of the relationship between word and genre remains the same, irrespective of the size of the data set. The chi-squared test also does not show the direction of any relationship. For example, if the relationship between the word grandma and speaker gender produces a chi-squared value of 35.5, we can be confident that these two variables are indeed related, but is grandma indicative of male or female speech? For 2×2 contingency tables, Yule's Distinctiveness Coefficient (DC), also known as Yule's Q, is a useful measure of both strength and direction of relationship. This measure will be described further in Section 10.3. For $2 \times n$ contingency tables, where either the number of rows or columns is more than 2, the strength of relationship can be found by Φ^2, where we divide the chi-squared value by the sample size. If both the number of rows and the number of columns are more than 2, we take the smaller of the number of rows and columns, and then subtract one. Φ^2 should be divided by this number to yield a measure called Cramer's V^2 (Lucy 2005: 48–51).

10.2.1 Comparisons Between British and American English

Hofland and Johansson (1982: 32) used the chi-squared technique to examine differences in the vocabulary used in American English and British English. They found, among other things, that British English prefers the –t form of the past

Table 10.3 Complete and absolute associations in LOB and Brown

(a) A complete association		
	Brown	*LOB*
theatre	0	63
theater	95	30
(b) An absolute association		
	Brown	*LOB*
south-west	0	10
southwest	16	0

tense, as in learnt and dreamt as opposed to learned and dreamed. Here it is useful to note Lucy's (2005: 49) distinction of 'complete' and 'absolute' associations between variables. In the case of British and American English, these associations could be between the choice of alternative word forms and the choice of corpus. For example, the prevalence of the two forms theatre and theatre can be tabulated as in Table 10.3(a). The association is said to be complete because there is a 0 in one of the cells, in this case denoting that theatre never occurred in the sample of American English (although theater sometimes occurs in the LOB corpus of British English). In an absolute association there is a 0 in both the cells on one diagonal. Table 10.3(b) shows the absolute association between the choice of south-west and southwest and the type of corpus. The former is never used in Brown, and the latter is never used in LOB. Complete and absolute associations can only be identified in 2 by 2 tables.

Leech and Fallon (1992), who also used the chi-squared test, regarded that differences in the relative frequency with which words were used in the United States and Britain could be indicative of cultural differences between the two countries. An obvious example of this would be the prevalence of baseball in Brown, and the prevalence of cricket in LOB.

10.2.2 Social Differentiation in the Use of Vocabulary

Rayson et al. (1997) studied the demographically sampled and spoken English component of the British National Corpus. They used the chi-squared test to look for differences in word frequency with gender, age and socio-economic group. Some of their results are shown in Table 10.4. In each case, the 20 words with the highest chi-squared score are presented.

In general, males use more taboo words and number words, while women use more first person and third person feminine pronouns. In fact, Rayson, Leech and Hodges report that 'taboo vocabulary is highly characteristic along all three dimensions of gender, age and social group', that is, among male speakers below 35 years of age in the C2/D/E social range. Such findings suggest that it is

Table 10.4 Vocabulary used predominantly by certain social groups

	A	*B*
Gender: Male (A) vs Female (B)	fucking, er, the, yeah, aye, right, hundred, fuck, is, of, two, three, a, four, ah, no, number, quid, one, mate	she, her, said, n't, I, and, to, cos, oh, christmas, thought, lovely, nice, mm, had, did, going, because, him, really
Age: Under 35 (A) vs Over 35 (B)	mum, fucking, my, mummy, like, wan 'na, goes, shit, dad, daddy, me, what, fuck, really, okay, cos, just, why	yes, well, mm, er, they, said, says, were, the, of, and, to, mean, he, but, perhaps, that, see, had
Socio-economic group: ABC1 (A) vs C2DE (B)[1]	yes, really, okay, are, actually, just, good, you, erm, right, school, think, need, your, basically, guy, sorry, hold, difficult, wicked, rice, class	he, says, said, fucking, ain't, yeah, its, them, aye, she, bloody, pound, I, hundred, well, n't, mummy, that, they, him, were, four, bloke, five, thousand

important to take into account the fact that variation may be the result of different factors acting in combination (e.g. the high prevalence of lovely in female speech is more due to the fact that it is older female speakers who use this word, rather than all females). Multivariate techniques are discussed in more detail in Section 10.3.

10.3 Genre Analysis

As described in Section 10.1, the LOB and Brown corpora were sampled according to text genre. The chi-squared test, as described in Section 10.2, can also be used to compare the vocabulary used in different genres. However, in this section, we will look at four other techniques for examining vocabulary differences in different genre, namely Yule's DC, hierarchical clustering, factor analysis and the SVM.

10.3.1 Yule's Distinctiveness Coefficient

Hofland and Johansson (1982, chapter 7) recorded the word frequencies in the following 'super-category' groups created by combining certain genres in the LOB corpus: A-C (newspaper text), D-H (informative prose), J (learned and scientific English), and K-R (fiction).The words for each super-category were ranked by their DC, which is a measure of the degree of over- or under-representation of a word in each category group. They give the example of the word spokesman which occurs 19 times (the 'absolute frequency') in one category, which we will call X. Taking into account the absolute frequency and total number of words in category

X, there is a relative frequency for spokesman of 107 words per million, using the formula:

> relative frequency = absolute frequency × 1 million/number of words in the category.

The word spokesman occurs 20 times in the million-word corpus as a whole (the corpus frequency). If the relative frequency is greater than the corpus frequency then the word is over-represented in that category, but if the relative frequency is less than the corpus frequency then the word is under-represented in that category. Since the relative frequency is 107 and the corpus frequency is only 20, spokesman is over-represented in category X by 87 words per million. The most over-represented words in category J and super-category K-R found by Hofland and Johansson are reproduced in Table 10.5. The words are listed in order of their DCs, highest first.

The 'distinctiveness coefficient' was originally developed by Yule (1944), and has since found acceptance outside linguistics. It appears, for example, in a book on forensic statistics (Lucy 2005: 48), where it is referred to as Yule's Q. It ties in with the ideas of complete and absolute associations described in Section 10.1, since in both cases Q is either –1 or 1. The form used by Hofland and Johansson is as follows:

$$DC = \frac{Freq_{LOB} - Freq_{BROWN}}{Freq_{LOB} + Freq_{BROWN}}$$

The DC always falls in the range –1 (for words exclusive to the Brown corpus) to +1 (for words exclusive to the LOB corpus). All other words have a DC somewhere in

Table 10.5 Most over-represented words in categories J vs K-R of the LOB corpus, listed in descending order of their DCs

Grammatical category	*J (science)*	*K-R (fiction)*
Nouns	constants, axis, equations, oxides, equation, theorem	mister, sofa, wallet, cheek, living-room, café
Lexical verbs	measured, assuming, calculated, occurs, assigned, emphasized	kissed, heaved, leaned, glanced, smiled, hesitated
Adjectives	thermal, linear, radioactive, structural, finite	damned, asleep, sorry, gay, miserable, dear
Adverbs	theoretically, significantly, approximately, hence, relatively, respectively	impatiently, softly, hastily, nervously, upstairs, faintly

Table 10.6 Comparison between the parts of speech of the words with highest distinctiveness ratio for two genres

Grammatical category	*J (science)*	*K-R (fiction)*
Nouns	58	23
Lexical verbs	1	31
Adjectives	12	2
Adverbs	0	4
Others	29	40
Total	**100**	**100**

between these two extremes. For example, the word abandoned is found 43 times in LOB and 25 times in Brown. Yule's DC is therefore $(43 - 25)/(43 + 25) = 18/68 = 0.26$. This positive value indicates that the word is over-represented in the LOB corpus, while a negative value would have shown that the word was under-represented in the LOB corpus (or over-represented in the Brown corpus).

Hofland and Johansson (1982: 32) made a comparison between the parts-of-speech of the 100 most over-represented words in both the J and K-R categories as determined by the distinctiveness ratio. In Table 10.6 there are more nouns and adjectives which are distinctive of J (learned and scientific articles) than for K-R (fiction), while there are more lexical verbs which are distinctive for fiction than for learned and scientific articles.

10.3.2 Hierarchical Clustering

The statistical techniques looked at so far are called univariate because a single variable is measured for each sampling unit. In the examples given in Sections 10.2 and 10.3.1, the single variable was frequency count, and the sampling units were single words or other countable linguistic feature. Sections 10.3.2 to 10.3.4 will examine multivariate approaches, where many different variables are all measured on the same sampling units. This section will focus on examples where a single genre (the sampling unit) is represented by a whole list of characteristic linguistic features, such as the frequency counts for each of 50 frequent words. The first multivariate technique examined will be a form of cluster analysis, which is a type of automatic categorization – similar things (such as related genres) are grouped together, and dissimilar things are kept apart.

10.3.2.1 Correlation in word-frequency rankings

The starting point for many clustering algorithms, such as the one described in the following section, is the similarity matrix, which is a square table of numeric scores reflecting how much each of the items (such as texts) to be clustered have in common with each of the others. The production of such a matrix is described

by Hofland and Johansson (1982: 19). To study the relationships between four corpora (LOB, Brown, Carroll et al.[2] (1971) and Jones and Sinclair[3] (1974)), they calculated the Spearman rank correlations (see below) for the frequencies of the 89 most common words in the LOB corpus. Common single words can often be indicators of grammatical and stylistic differences between genres. For example by indicates use of the passive voice, which indicates the use of relative clauses, the indicates the use of nouns, and an the use of Latinate vocabulary (ibid.: 22). To illustrate their method of producing the similarity matrix for the four corpora, it is repeated here for just the ten most frequent words in the LOB corpus. These ten words and their ranks in each of the corpora (where the most frequent word would have a rank of 1) are listed in Table 10.7(a). This data is rewritten in Table 10.7(b),

Table 10.7 The ten most frequent words in the LOB corpus compared with the ranks of the corresponding words in three other corpora

(a) Considering all words

	LOB	*Brown*	*Carroll et al. (1971)*	*Jones and Sinclair (1974)*
the	1	1	1	1
of	2	2	2	8
and	3	3	3	3
to	4	4	5	6
a	5	5	4	5
in	6	6	6	9
that	7	7	9	11
is	8	8	7	12
was	9	9	13	14
it	10	12	10	7

(b) Considering only the top ten words in the LOB corpus

	LOB	*Brown*	*Carroll et al. (1971)*	*Jones and Sinclair (1974)*
the	1	1	1	1
of	2	2	2	6
and	3	3	3	2
to	4	4	5	4
a	5	5	4	3
in	6	6	6	7
that	7	7	8	8
is	8	8	7	9
was	9	9	10	10
it	10	10	9	5

except that the words that were ranked tenth and eleventh in the Brown corpus are ignored, and only the ranks of each word among the top ten words in LOB are considered. Similarly, was is recorded in Table 10.7(a) as being the 13th most common word overall in the Carroll et al. corpus. Ignoring the words that were 11th and 12th most common in Carroll et al., since these words are not among the top ten10 in LOB, we record in Table 10.7(b) that of the words which appear in the top ten of LOB; was was the 10th most frequent in the Carroll et al. corpus.

A correlation coefficient measures the similarity between two corresponding sets of data. One such measure is the Spearman rank correlation coefficient (see, e.g. Woods et al. 1986: 170–1), which is given by the following formula:

$$r_s = 1 - \frac{6 \times \sum (R - S)^2}{n(n^2 - 1)}$$

When comparing the word frequency rankings for LOB and Jones and Sinclair, the top line of this formula can be calculated by taking the following steps. In Table 10.8 R is the rank of a word in LOB and S is the rank of the same word in Jones and Sinclair. In column 3 the difference in the ranks is recorded, and in column 4, the squares of the differences between the ranks is calculated. The summation symbol Σ in the bottom row means that the sums of the squares of the differences between the ranks must be added together, giving a total of 50. Finally, n is the number of words considered in the analysis, which is 10.

Spearman's rank correlation coefficient is then:

Table 10.8 Calculation of the top line of the Spearman formula (comparing LOB and Jones and Sinclair)

	R = Rank (LOB)	*S = Rank (Jones and Sinclair)*	*R – S*	*$(R - S)^2$*
the	1	1	0	0
of	2	6	–4	16
and	3	2	1	1
to	4	4	0	0
a	5	3	2	4
in	6	7	–1	1
that	7	8	–1	1
is	8	9	–1	1
was	9	10	–1	1
it	10	5	5	25
				Σ = Total = 50

$$1-\frac{6\times 50}{10\times 99}=0.697$$

A correlation coefficient is given as a number between –1 and +1. A coefficient of +1 means that two sets of scores have perfect positive correlation, whereas –1 indicates perfect negative correlation (as one score increases, the other decreases). Scores closer to zero indicate weaker correlation while a score of zero means that there is no correlation at all between the two sets of scores.

Once all of the corpora have been compared against each other in this way the similarity matrix in Table 10.9 can be obtained. The trivial observation that the word frequency ranking for a corpus is identical with itself does not need to be recorded, so the principal diagonal from top left to bottom right can be left empty. Note also that the matrix is symmetrical, so for example the similarity between Brown and Carroll is the same as that between Carroll and Brown. Since the rankings for LOB and Brown were identical, the Spearman correlation between them is +1. Hofland and Johansson also found that although Brown and Carroll contain American English, and the other two corpora contain British English, the most similar pair of corpora based on the word frequency rankings were LOB and Brown. It could be concluded that this is because these corpora were sampled to match each other with respect to the number of texts of each type they contain.

The similarity matrix of Table 10.9 can also be calculated using the Matlab statistical toolbox. The input data is that of Table 10.7(b), in a form where each corpus is allocated a row rather than a column:

```
Data = [ 1  2  3  4  5  6  7  8  9 10;
         1  2  3  4  5  6  7  8  9 10;
         1  2  3  5  4  6  8  7 10  9;
         1  6  2  4  3  7  8  9 10  5 ]
```

The command `M = pdist(Data,'spearman')` then produces a difference matrix called M, showing the differences between the corpora rather than their similarities. To transform this into a similarity matrix called N, we need the

Table 10.9 Similarity matrix for four corpora based on the word frequency rankings for ten common words

	LOB	*Brown*	*Carroll et al.*	*Jones and Sinclair*
LOB	–	1.000	0.9636	0.6970
Brown	1.000	–	0.9636	0.6970
Carroll et al.	0.9636	0.6970	–	0.7576
Jones and Sinclair	0.6970	0.6970	0.7576	–

command `N = 1 - M`. When this matrix is printed out it consists of the six numbers in the top right-hand corner of Table 10.9. Since Table 10.9 is symmetrical, it is not necessary to reproduce the data for both corners. Thus Matlab gives the matrix in the following form:

```
N = [1.000 0.9636 0.6970 0.9636 0.6970 0.7576]
```

10.3.2.2 Production of a dendrogram

This section describes how clustering procedures are used in order to show relationships between groups of data (in this case the 15 different genres in LOB). Clustering is a technique which starts by considering all of the groups separately and calculating which two are the most similar. Then the two most similar groups are joined together to form a single cluster and the similarity calculation is carried out again, with the next two most similar groups being joined together, and so on, until there is only one 'super-cluster' left. The process can be visually represented as a dendogram – a type of tree diagram which shows the relationships between the groups.

Starting with the similarity matrix of genres produced by Hofland and Johansson (1982: 23), based on the rankings of the most common 89 words in LOB in each genre, these are rewritten in the linear format of matrix N as shown in the previous section. The matrix is then transformed into a distance matrix M with M = 1 – N. The Matlab statistical tool box can then be used to cluster the genres and produce the dendogram in Fig. 10.1. The letters along the bottom of Fig. 10.1 refer to the various genres in LOB, while the numbers down the side axis are difference scores – the lower the score here, the more similar the genres are. Therefore, the most similar pair of genres were K (fiction) and P (romance) which have a similarity of 0.97 or a difference of 1 – 0.97 = 0.03. The first step then, is to join these two genres into a single cluster. Henceforth, these two genres are considered as a singe entity.

The next most similar pair of nodes are L (mystery and detective fiction) and N adventure and western fiction), with a difference of 0.04, so these two are now fused to form a single entity. At the third stage, the two most similar entities are not individual nodes, but the clusters formed at the previous two stages. The distance between two clusters can be calculated in various ways. Here 'average linkage' was used, where the distance between two clusters is the average difference of all the members of one cluster and all the members of the other cluster. When using the 'single linkage method', the distance between two clusters is taken as the distance between the most similar member of one cluster to the most similar member of the other. The two clusters created so far are now joined together to form a larger cluster of four nodes. At each subsequent stage of the process the two most similar entities (clusters or individual nodes) are fused, until a single cluster remains. Matlab conducts the clustering process, described here in response to the command `Z = linkage(M,'average')`, where N is the name of the original difference matrix.

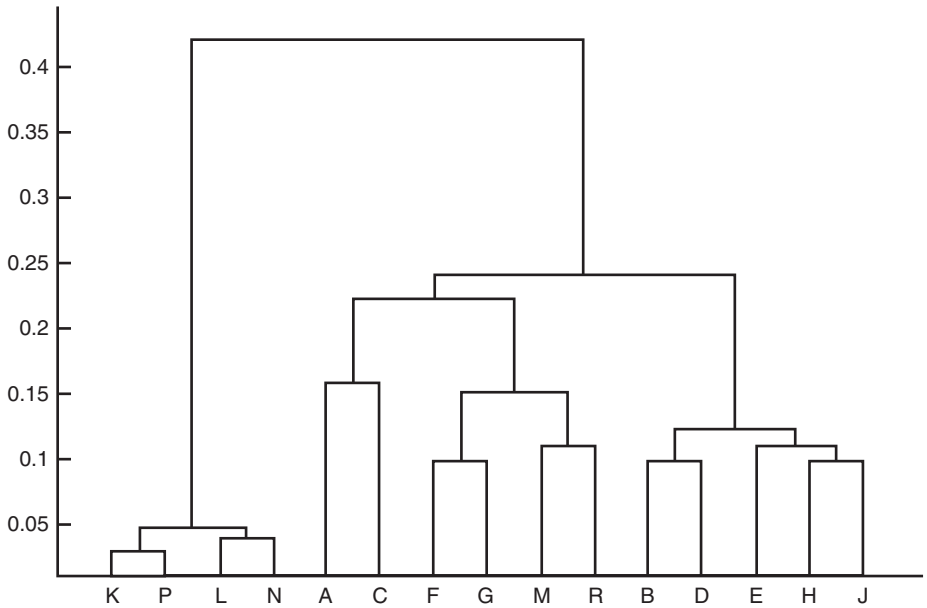

Figure 10.1 Dendrogram for the Genres in the LOB corpus.

The resulting dendogram can be plotted out on Matlab using the command `H = dendrogram(Z)`. The clusters are normally labelled numerically (in this case 1 to 15), but in order to allow the genres to be recognized according to the alphabetic labels in Table 10.1, it was necessary to input a couple of additional commands into Matlab in order to create the matrix `V = [A; B; C; D; E; F; G; H; J; K; L; M; N; P; R]`, and then `H = dendrogram(Z,'labels',V)`.

In Fig. 10.1, Genres K (General fiction), P (Romance and love story), L (Mystery and detective fiction) and N (Adventure and western fiction) form the most tightly bound cluster – not only were they closely related to each other, but they were quite distinct from all the other genres. It is not surprising that four types of fiction should be closely related to each other. If a criterion is set that all genres with less than 0.3 difference to each other belong in the same cluster, then there are just two clusters: the cluster containing K, L, N and P vs the rest. However, if a stricter criterion is set, that only genres with less than 0.2 difference should be considered as members of a common cluster, then the data set splits into four clusters. As before, the first of these is the cluster of K, L, N and P. After that, reading from left to right, a second cluster consists of two types of newspaper material: A (Press reportage) and C (Press reviews). It seems that these two genres are more intimately related to each other than they are to B (Press editorial), which appears in the fourth cluster (but remember that the second, third and fourth clusters are not so distantly removed from each other as they are from the first cluster). The third cluster suggests that M (Science fiction) and R (Humour) have much more

in common with each other, along with F (Popular lore) and G (Belles lettres) than they do with the other types of fiction in the first cluster. Finally, the fourth cluster consists of B (Press editorial), D (Religion), E (Skills, trades and hobbies), H (Government documents), and J (Learned and scientific writings). Perhaps these five genres all aim at a common 'factual' writing style.

It should be noted again that the criteria for comparison in Fig. 10.1 were based on the rankings of the most common 89 words in each genre. It is possible that other types of criteria (e.g. standardized type-token ratio, overall proportion of nouns, mean word length) might produce different clustering patterns.

Although Hofland and Johansson used a visual representation of the similarity matrix, where dark coloured squares indicated very similar genres and lighter squares indicated less similar genres (ibid.: 24–5), their findings were similar to those shown in the dendrogram: the corpus was seen to have two major groups of texts (informative and imaginative prose), bridged by categories of what they called 'essayistic prose' – popular lore, belles lettres, science fiction and humour (ibid.: 27).

10.3.3 Factor Analysis

Biber (1988) writes that texts can be quantitatively characterized by any countable features such as single word frequency counts. Biber also used other countable features such as suasive verbs (agree, arrange, ask, beg, command, etc.), contractions, and the type-token ratio. In the technique of factor analysis however, text types are characterized not so much by individual markers, but by their regular co-occurrences within these text types. Groups of features which consistently co-occur in texts, using Biber's terminology, are said to define a 'linguistic dimension'. Such features are said to have positive loadings with respect to that dimension, but it is also possible for dimensions to be defined by features which are negatively correlated with them, that is, negative loadings. Biber's approach is to group the features first, and then to identify the linguistic dimensions, such as formal vs informal, or involved vs detached. The relative importance of factors is measured by values called eigenvalues. It is something of an art to empirically assess how many dimensions there really are in a set of texts, but one technique is to manually examine a graph of eigenvalues called a scree plot. A one-dimensonal example, showing a single linguistic dimension is given by Biber (1988: 17). At one pole is 'many pronouns and contractions', near which lie conversational texts, and almost as close, panel discussions. At the opposite pole, 'few pronouns and contractions', are scientific texts and fiction. Biber (1988: 63) gives the following methodological overview of factor analysis.

1. The use of computer corpora, such as LOB, which classify texts by a wide range of genres.
2. The use of computer programs to count the frequencies of linguistic features throughout the range of genres. In the experiment described below, Perl scripts

were written in order to count the occurrences of the top 50 words in the LOB corpus.
3. Use of factor analysis, to determine co-occurrence relations among the features.
4. Use of the linguist's intuition to interpret the linguistic dimensions discovered.

In a small experiment, the author wrote a Perl script to count the raw frequencies within each genre of the 50 most common words in the LOB corpus. These data were stored in a matrix called B, where the columns referred to genre and the rows referred to linguistic feature (word). Following Biber (1998: 94), these raw frequency data were standardized to a mean of 0 and a standard deviation of 1, using the Matlab command `[S, meanp, stdp] = prestd(p)`, which creates a matrix of standardized data called S. The use of standardized data prevents those features that occur very frequently (such as the word the) from having an undue influence on the computed factor scores. The factor analysis itself uses a process called 'rotation', which simplifies the analysis by ensuring that each feature loads on as few factors as possible. Thus each factor becomes characterized only by its most representative features. One type of rotation called 'varimax' assumes the factors are uncorrelated, while another called 'promax' permits factors to overlap in features. Matlab uses varimax for the calculation of a matrix of factor loadings (called lambda) and promax for producing a convenient visual representation of the data called a biplot. The feature loadings in lambda should all be in the range –1 to +1. The factor analysis itself takes our matrix of standardized frequency counts S, with the command:

```
[lambda, psi, T, stats, F] = factoran(S, 2)
```

This will give a solution with two factors, and ignore all other factors with smaller eigenvalues than these. The number '2' can be replaced with higher numbers if desired, in order to enable the extraction of more factors. The value 2 was simply chosen here to produce the clearest biplot for illustration purposes, and the author did not attempt to calculate the optimal number of factors from a mathematical point of view. As stated above, lambda is a matrix of factor loadings. F is the matrix of scores for each of the features, indicating which words are associated with which factors.

Having performed the factor analysis, the next step is to produce a visual representation called a 'biplot' which shows the relationships between the factors, the text genres, and the individual linguistic features (in this case, frequent words). The biplot in Fig. 10.2 was produced using the following Matlab command:

```
biplot(lambda, 'scores', F, 'varlabels', V)
```

Here, `'scores'` is a command used to plot the individual frequently occurring words on the biplot as red dots. V is a one-dimensional matrix of genre labels created beforehand, using the command `V = [A;B;C;D;E;F;G;H;J;K;L;M;`

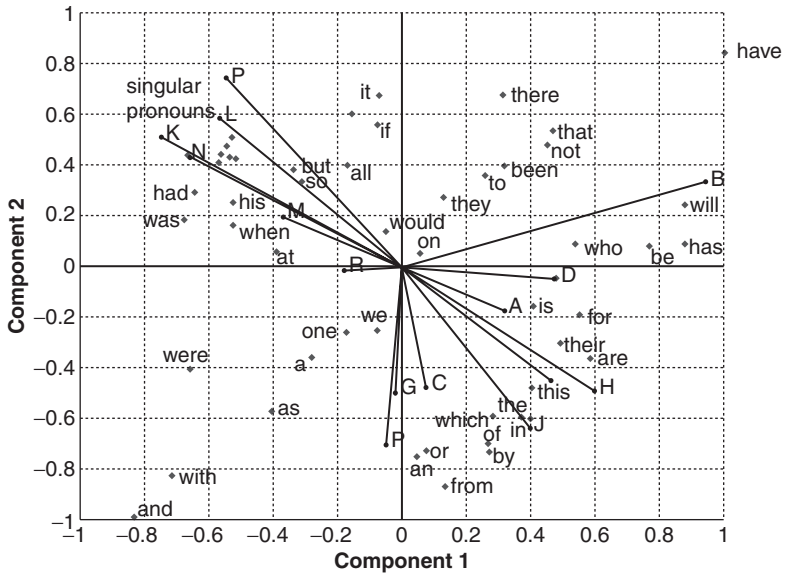

Figure 10.2 Biplot for a factor analysis.

N;P;R]. The term 'varlabels' is the command used to print these labels next to the appropriate text genre vectors (in blue). The scores are scaled by the biplot, so the word points all fit inside the unit square. In Fig. 10.2, the word points have been labelled manually. As was the case for the dendrogram shown in Fig. 10.1, it can be seen that the four types of fiction K (General), L (Mystery and detective), N (Adventure and Western) and P (Romance and love story) are closely related to each other, scoring highly on factor 2 but loaded highly and negatively on factor 1. The genres M (Science fiction) and R (Humour) are again closely related. Like the other fictional texts, they are loaded positively on factor 2 and negatively on factor 1, but with much smaller loadings. B (Press: Editorial) is unique in that it is the only genre to be positively loaded on both factors, but as with the dendrogram, its closest neighbour is D (religion). The genres F (Popular lore), G (Belles lettres) and C (Press: Reviews) are again closely related, this time all independent of factor 1 and negatively loaded on factor 2. The genres D (Religion), E (Skills, trades and hobbies), H (Misc) and J (Science) were also close neighbours in both analyses, here negatively loaded on factor 2 but positively loaded on factor 1. Only genre A (Press: Reportage) is somewhat differently located with respect to its neighbours compared with the dendrogram. Table 10.10 shows the words most closely associated (positively and negatively) with factors 1 and 2.

The most striking features of Table 10.10 are:

(a) the word have had the most positive score with respect to both factors,
(b) the word and had the most negative score with respect to both factors, and

Table 10.10 Scores for factors 1 and 2

Factor 1				*Factor 2*			
Positive loadings		*Negative loadings*		*Positive loadings*		*Negative loadings*	
word	*score*	*word*	*score*	*word*	*score*	*word*	*score*
have	1.00	and	–0.83	have	0.83	and	–0.99
will	0.88	with	–0.72	it	0.68	from	–0.87
has	0.88	was	–0.68	no	0.60	with	–0.83
be	0.77	were	–0.65	if	0.56	an	–0.75
are	0.58	had	–0.64	that	0.53	by	–0.74
who	0.54	you	–0.57	she	0.50	or	–0.73
their	0.48	him	–0.56	her	0.48	of	–0.70
that	0.47	I	–0.54	not	0.45	in	–0.61
not	0.45	she	–0.53	him	0.44	the	–0.60
this	0.40	when	–0.53	I	0.43	which	–0.59
in	0.40	he	–0.52	he	0.34	as	–0.58
		his	–0.52			this	–0.48
		as	–0.40			were	–0.41

(c) the singular personal pronouns (I, you, he, she, his, him, her) were clustered very closely together.

In Fig. 10.2 the six points between the lines for genres K (General fiction) and L (Mystery and detective fiction) correspond to she, her, him, he, I and you. This shows that singular personal pronouns are very characteristic of fictional texts.

10.3.4 Linguistic Facets and SVMs

Santini (2006) developed an alternative approach to the automatic classification of web texts based on genre. Some web genres are 'extant', in that they have long existed in other media, such as the printed word. Examples of these are found in the BBC Web Genre Collection, namely editorials, do-it-yourself, mini-guides, short biographies and features articles. 'Variant' genres, on the other hand, have only arisen since the advent of the world-wide web, such as those found in the 7-Web-Genre collection: blogs, e-shops, frequently-asked-questions, listings (such as site maps or indexes), personal home pages and search pages (which include dynamic content, especially some form of search engine). The 7-Web-Genre collection also contains an extant genre: that of online newspaper front pages. It is anticipated that new web genres will emerge in the not too distant future. Web pages often belong to more than one genre, while others do not fit into any genre

at all. Thus Santini's automatic genre identifier allows zero, one or more than one genre to be assigned to a given input text. As was the case with Biber's factor analysis, the first stage in genre analysis involves counting automatically extractable genre-revealing features. These include the identification of such features as function words and punctuation marks, which require no pre-processing by the computer. Other, more linguistically sophisticated features require the use of a tagger and a parser, such as part-of-speech trigrams. The third group of genre-identifying features are called 'facets', sets of features sharing a common textual interpretation, such as 'first person facet', which include all first person pronouns. Other facets consist of genre-specific referential vocabulary, such as the set '£, basket, buy, cart, catalogue, checkouts, cost . . .', all of which suggest the 'e-shop' genre, and facets based on the presence of HTML tags in the text. For example, the 'functionality facet' includes tags such as '<button>' and '<form>' which indicate the presence of buttons and forms in an interactive web page.

Having identified a range of countable features for describing texts, the next step is to choose a classifier capable of discriminating between sets of features in a text to predict the most likely genre or genres. The selected classifier is a SVM, which is a form of supervised learner. This means that the classifier has to be presented with an adequate number of training examples, consisting both of the numbers of each feature found in a test, and the genre or genres a knowledgeable human has assigned to them. Once the classifier is able to discriminate between sets of features belonging to different genres, a new phase begins, called the test phase. Now the feature sets of previously unseen texts are presented to the classifier, and the classifier alone assigns genres to them. This approach contrasts with hierarchical clustering and factor analysis, which are unsupervised learning approaches. In unsupervised learning, at no stage is any human annotator required to classify any examples. The classifier is able to learn everything it needs to know from the unannotated texts presented to it. The objective of this work was not to make new linguistic discoveries, although questions such as 'what constitutes an e-shop genre?' will have been partially answered by the qualitative analysis required in compiling the linguistic facets. The main finding was that automatic genre analysis has practical uses, in this case helping to realize the goal of genre-sensitive search engines, which will answer queries such as 'Find me an academic article about X'.

10.5 Computational Stylometry

Up to now, differences between types of text have been considered, with each text type having been produced by many different authors or speakers. In this section, I briefly discuss how some of the techniques encountered in this chapter have been used to examine the writing styles of individual authors (see also Malhberg, this volume).

The chi-squared test has been used by Forsyth (1999) to compare the poetry writing styles of the younger and the elder Yeats. His choice of linguistic feature

Table 10.11 Top six discriminators for the younger and older Yeats

Rank	*Substring*	*Chi-squared*	*YY-count*	*OY-count*
1	what	35.1	30	100
2	can	34.3	21	82
3	s, an	25.4	63	19
4	whi	25.4	67	21
5	with	22.3	139	74
6	?	21.9	30	83

was the 'substring', any sequence of up to eight characters, including spaces and punctuation, which appeared in a set of Yeats' poems. Since there are so many of these, a random selection was taken. Having determined which substrings were most typical of the younger and the elder Yeats, Forsyth was able to estimate the dates of other Yeats' poems by how many 'older' and 'younger' features they contained, using a 'Youthful Yeatsian Index' measure similar to Yule's DC. The substrings most characteristic of the younger and of the older Yeats are shown in Table 10.11. Although these substrings were randomly selected, they do reveal some higher level linguistic constructs. Both what and ? indicate that the poems of the older Yeats contain more questions.

Hierarchical clustering techniques were effectively used by Holmes (1992) in a comparison of Mormon scripture, Joseph Smith's personal writings and sections of the Old Testament. The main finding was that although these three sources could easily be distinguished from one another, the various prophets who had supposedly written the Mormon books over a period of many centuries could not be distinguished from each other. Holmes et al. (2001) used Principal Components Analysis, a technique closely related to factor analysis, to determine the provenance of a collection of letters, purportedly by an American War General, George Pickett, and published in a book called *The Heart of a Soldier*. They compared these letters with the autobiography of LaSalle Pickett, George's widow, George Pickett's personal pre- and post-war letters, George Pickett's war reports, genuine handwritten letters by George Pickett (the Inman Papers), and Walter Harrison's book *Pickett's Men*. Each set of texts was internally consistent, forming close clusters on the chart. However, the texts from *The Heart of a Soldier* also lay very close to LaSalle Pickett's autobiography, suggesting that she may have been the true author of the letters supposedly written by her husband. Finally, Popescu and Dinu (2007) used a SVM to discriminate between the possible authors of the disputed 'Federalist Papers'. These papers, written in 1787–1788 to encourage the American people to accept the new constitution, and published in various newspapers under the pseudonym 'Publius', were either written by Alexander Hamilton or James Madison. They are an ideal test bed for trying out new techniques in computational stylometry, because although the writing styles of these two writers were very similar, there

are only two of them to choose between. The SVM indicated that Madison was the more likely author in all 12 cases, a finding which is in accordance with previous computer analyses and most historians.

10.6 Discussion

This chapter has examined various statistical methods for distinguishing between different text types, arising through such factors as demographic variation in writers or speakers, genre differences, topic differences or individual writing styles. I wish to end by discussing some of the potential issues that need to be taken into account when using these statistical techniques to examine variation in corpora.

As mentioned previously, the main difficulty with these approaches is the fact that each of these sources of linguistic variation can obscure any of the others. One way that authors have tried to compensate for this is by using texts that are as similar as possible. For example, in their study of the frequencies of syntactic structures used by different authors, Baayen, van Halteren and Tweedie (1996) used only science fiction texts, so the individual stylistic differences they were looking for did not get swamped by genre differences.

When carrying out comparisons, however, it is often difficult to build corpora that only differ on a single factor. For example, although little has been said here about cultural corpora, the main deviation from matching the LOB and Brown corpora text-for-text as closely as possible was due to cultural factors. So more extracts from westerns are found in category N (Adventure and western fiction) of the Brown corpus, since more westerns are written in the United States (Hofland and Johansson 1982: 3). Another issue arises regarding whether a diachronic corpus be exactly balanced text-for-text, or whether the diachronic corpus should take into account actual cultural differences. In using the same sampling model for the Brown family of corpora, the corpus builders chose the first option. However, it could be argued that perhaps the sampling model should have been changed over time in order to reflect changing patterns of text production and reception. For example, the 1960s could be seen as the heyday of science fiction, with more people writing and buying science fiction than in later decades. Perhaps the composition of a diachronic corpus should take this into account, although this would strongly impact on the amount and type of variation within the corpus.

In addition, it needs to be borne in mind that individual authors do not have a homogenous writing style, even when writing in a single genre. For example, DeForest and Johnson (2001) were able to show linguistic variation within a single author, by finding the proportion of Latinate vocabulary used by various characters in Jane Austen's novels. They found that a high proportion of Latinate vocabulary was indicative of, among other things, education, maleness and euphemism. A corpus with a small number of authors may result in variation that is due to individual writing styles. However, the more authors we include in a corpus (with the aim of averaging out individual differences) the more likely it is to introduce yet more factors that have to be balanced for. So if we build two corpora (say from

culture a and culture b or from time period a and time period b) each containing equal-sized samples of writing from say, 500 authors, we might also need to take into account their age, gender, social class, education level and the like, otherwise imbalances in any of those factors could impact on any variation found.

Anecdotally, differences in subject matter are said to mask other types of linguistic variation. However, the linguistic features most suitable for differentiating texts according to subject matter are mid-frequency terms, while other types of linguistic features have proved effective at telling apart genres (such as the high frequency words used in this chapter, and Santini's linguistic facets), chronology (Forsyth's substrings) and individual authors (function words said to be outside the author's conscious control). All this suggests that the problem of different types of linguistic variation masking each other can be alleviated by finding linguistic features which are particularly good at identifying one source of linguistic variation, without being indicative of others. Such a proposal suggests a way forward for a more robust form of studies of corpus variation.

Notes

1 ABC1 and C2DE refer to the British government's classification of socio-economic status. Group A (3% of the population) represents professional people, senior managers or top level civil servants. Group B (20%), are middle-management executives, owners of small businesses or principle officers in local government. Group C1 (28%) are people in junior managerial positions and everyone else in non-manual positions. Group C2 (21%) are skilled manual workers. Group D (18%) are semi-skilled and unskilled manual workers. Group E (10%) are people on state benefit, the unemployed or casual workers. See *Occupation Groupings: A Job Dictionary* (2006), The Market Research Society.

2 The corpus reported in Carroll et al. (1971) consists of 5 million words of written American English used in schools over a range of subject areas and grades of writing.

3 The Edinburgh-Birmingham corpus used by Jones and Sinclair (1974) consists of collections of transcribed British speech from the 1960s.