Using electronic corpora in historical dialectology research : the problem of document length variation

Hermann Moisl
University of Newcastle upon Tyne

Introduction

The proliferation of computational technology has generated an explosive production of electronically encoded information of all kinds. In the face of this, traditional philological methods for search and interpretation of data have been overwhelmed by volume, and a variety of computational methods have been developed in an attempt to make the deluge tractable. These developments have clear implications for corpus-based linguistics in general, and for corpus-based study of historical dialectology in particular: as more and larger historical text corpora become available, effective analysis of them will increasingly be tractable only by adapting the interpretative methods developed by the statistical (Hair *et al*. 2005; Tabachnik & Fidell 2006), information retrieval (Belew 2000; Grossman & Frieder 2004), pattern recognition (Bishop 2006), and related communities. To use such analytical methods effectively, however, issues that arise with respect to the abstraction of data from corpora have to be understood. This paper addresses an issue that has a fundamental bearing on the validity of analytical results based on such data: variation in document length. The discussion is in four main parts. The first part shows how a particular class of computational methods, exploratory multivariate analysis, can be used in historical dialectology research, the second explains why variation in document length can be a problem in such analysis, the third proposes document length normalization as a solution to that problem, and the fourth points out some difficulties associated with document length normalization

1. Exploratory multivariate analysis in historical dialectology

Historical dialectology is based on the study of collections of spoken or written language. A typical research question is: given a corpus comprising a collection of historical documents, can those documents be dialectally classified on the basis of their linguistic characteristics --phonetic, phonological, morphological, lexical, or syntactic? There are two main approaches to this type of question:

- Theoretically-driven: Classification criteria are selected by the researcher on the basis of an independently-specified theoretical linguistic framework supported by existing case studies conducted within that framework and by personal knowledge of the characteristics of the language in the historical period in question.

- Empirically-driven: Classification criteria are algorithmically abstracted from the corpus data itself without reference to any theoretical linguistic framework, existing analytical results, or personal knowledge of the subject domain.

The theoretically-driven approach is suitable where the corpus is embedded in a well understood dialectological context, while the empirically-driven one is suitable where little is known a priori about it. This paper is concerned with analysis of historical corpora whose characteristics are not well known using an empirically-driven methodology called exploratory multivariate analysis.

1.1 *The nature of exploratory multivariate analysis*

In current scientific practice, a hypothesis about some natural phenomenon is proposed and its adequacy assessed using data

obtained from observation of the domain of inquiry. But nature is dauntingly complex, and there is little practical or indeed theoretical hope of being able to observe even a small part of it exhaustively. Instead, the researcher selects particular aspects of the domain which seem salient to his or her research question. Each selected aspect is represented by a variable, and a series of observations is conducted in which, at each observation, the values of each variable are recorded. A body of data is thereby built up on the basis of which the hypothesis can be assessed. If only one aspect of the domain is observed --the height of individuals in a population, say-- then the data consists of some number of values assigned to a single variable; that data is univariate. If two aspects are observed --say height and weight-- then the data is bivariate, if three trivariate, and so on up to some arbitrary number *n*. Any data where *n* is greater than 1 is multivariate.

As the size of the data grows, that is, as the number of variables and / or the number of observations increases, it becomes ever more difficult to see any interesting regularities by direct inspection. Take, for example, data in which three persons p1, p2, and p3 are described in terms of two variables 'age' and 'weight'. If p1 is young, p2 middle-aged, and p3 old, and if p1's weight is low, p2's medium, and p3's high, it's easy enough to infer just by looking at the data that weight increases with age, at least for this sample. Adding a third variable 'height' makes it a little more difficult but not impossible to see such regularities. But what if there are a dozen variables, including such things as 'income', 'hair colour', and 'shoe size'? It is very difficult to see very much in the data, and, of course, there is no limit to the number of variables that might be used to describe people. How easy would it be to see any regularities in data with, say 100 variables, even for only three people? And, as the number of persons increases, so does the difficulty of interpretation. In short, as the number of variables and/or observations grows, so does the difficulty of conceptualizing the interrelationships of variables on the one hand, and the interrelationships of observations on the other.

Exploratory multivariate analysis is a general term for mathematically based methods for discovering and understanding data that has too many variables for it to be comprehensible via direct inspection (Andrienko & Andrienko 2005). Exploratory methods have long been used in numerous science and engineering disciplines as a

way of generating hypotheses about data whose characteristics are not well understood. Closer to home, the proliferation of electronic text in recent decades has seen the application of exploratory methods in processing of natural language text in areas like information retrieval (Belew 2000; Grossman & Frieder 2004) and data mining (Tan *et al*. 2006), as well as in linguistic analysis and in traditional philology more generally. The literature on linguistic and philological applications is too large and varied to cite here; representative dialectological examples are Heeringa & Nerbonne (2001), and Nerbonne & Herrings (2001).

1.2. *Application of exploratory multivariate analysis to historical dialectology*

Exploratory multivariate analysis methods are intended specifically to classify any given collection of objects described by more or less numerous variables. Because this is precisely the kind of research question with which historical dialectology is often concerned, their extension to corpus analysis is a natural step.

To exemplify this extension we consider the *Newcastle Electronic Corpus of Tyneside English* (NECTE), a corpus of dialect speech from Tyneside in North-East England (Allen *et al*. 2005). It includes phonetic transcriptions of 63 interviews together with social data about the speakers, and as such offers an opportunity to study the phonetic dialectology of Tyneside speech of the late 1960s. Moisl & Jones (2005), Moisl *et al*. (2006), Moisl & Maguire (2008) have begun that study using exploratory analysis of the transcriptions with the aim of generating hypotheses about phonetic variation among speakers and speaker groups in the corpus. These studies were based on comparison of profiles associated with each of the TLS speakers. A profile for any speaker S is the number of times S uses each of the phonetic segments in the NECTE transcription scheme in his or her interview. More specifically, the profile P associated with S is a vector having as many elements as there are segments such that each vector element P_j represents the *j*'th segment, where *j* is in the range 1..number of segments in the NECTE phonetic transcription scheme, and the value stored at P_j is an integer representing the number of

times S uses the *j*'th segment. There are 156 segments, and so a speaker profile is a length-156 vector. There are 63 TLS speakers, and their profiles are represented in a matrix N having 63 rows, one for each profile; a fragment is shown in Figure 1:

	v1: i	v2: i̞	...	v156: ɔ̃
Speaker 1	23	4	...	7
Speaker 2	3	56	...	4
...	...	...	...	...
Speaker 63	18	35	...	8

Figure 1: NECTE phonetic segment frequency data matrix N

The aim is to classify the 63 speakers in accordance with the values in their speaker profiles.

N is an example of a data that is simply too large and complex to be interpretable by direct inspection. It was therefore analyzed using hierarchical cluster analysis (Everitt *et al.* 2001), a widely used exploratory analytical method that represents relative similarity among items in high-dimensional data as a nested tree:

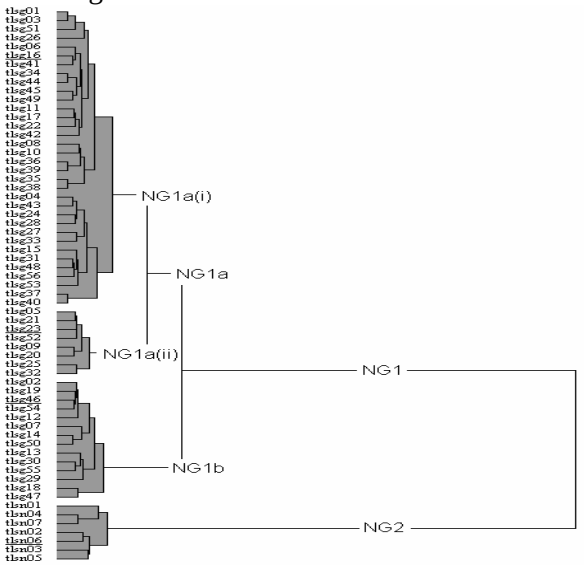


Figure 2: Cluster analysis of the NECTE data matrix N

Cluster trees like this are familiar to linguists as representations of syntactic phrase structure, but differ from linguistic trees in that

- they are shown horizontally rather in the vertical orientation that is more usual for linguistic phrase structure trees in order to make them more readily representable on a page.

- the leaves are not lexical tokens but labels for the data items --here speaker labels.

- they represent not grammatical constituency but relativities of similarity between the vectors representing the data entities --here speakers. The lengths of the branches linking the subtrees represent degrees of similarity: the shorter the branch, the more similar the subtrees. Thus the subtrees labelled NG1 and NG2 above are very dissimilar, NG1a(i) and NG1a(ii) very similar, and so on.

The hierarchical analysis partitions the NECTE speakers on the basis of their phonetic usage (Moisl *et al.*, 2006). The main distinction is between middle class, well educated speakers from Newcastle on the north side of the river Tyne, labelled NG2, and working class, less well educated speakers from Gateshead on the south side of the Tyne, labelled NG1. The Gateshead speakers are categorized into NG1b (exclusively male), and NG1a (mainly through not exclusively female); NG1a is subcategorized into NG1a(i) (working class females) and NG1a(ii) (males and females with relatively higher socioeconomic status).

2. The problem of variation in document length

It is a simple fact of life that documents in any given collection can vary considerably in length. Where the data abstracted from such a corpus is based on frequency, such length variation is a problem for cluster analysis. This section shows why.

For concreteness of exposition the discussion is based on a small, artificially-constructed corpus C with known structural characteristics. It comprises 9 excerpts from historical English texts from Old English to Early Modern English. These are arranged chronologically in Figure 3:

Name	**Date**	**Size**
Sermo Lupi ad Anglos	996 - 1023 AD	13 kb
Beowulf	c.1000 AD	106 kb
Apollonius of Tyre	c.1000-1050	35 kb
The Owl and the Nightingale	c.1250-1300 AD	10 kb
Chaucer, *Troilus & Criseyde*	c.1370 AD	123 kb
Malory, *Morte d'Arthur*	c.1470 AD	132 kb
Everyman	c.1500 AD	37 kb
Spenser, *Faerie Queene*	1590 AD	34 kb
King James Bible	1611 AD	11kb

Figure 3: The contents of example corpus C

2.1 *Data creation*

Prior to its standardization in the later 18th century, spelling in the British Isles varied considerably from time to time and place to place, reflecting on the one hand differences in phonetics, phonology and morphology at different stages of linguistic development, and on the other differences in spelling conventions. It should, therefore, be possible to categorize texts on the basis of their spelling, and to correlate the resulting categorizations with chronology. This, therefore, is the research question: can the documents in C be accurately categorized chronologically solely on the basis of their spelling?

How does one go about investigating spelling? The approach taken here is based on the concept of the tuple. A tuple is a sequence of symbols: AA is a pair, AAA a triple. AAA a four-tuple, and so on. This concept of tuple offers an efficient way of comparing spellings among texts:

- Given a collection D containing *m* documents, compile a list of all letter tuples of that occur in the texts. Assume that there are *n* such tuples.

- To each of the documents d_i in D (for $i = 1..m$) assign a vector of length *n* such that each vector element v_j (for $j = 1..n$) represents one of the *n* letter tuples.

- In each document d_i count the number of times each of the *n* letter tuples *j* occurs, and enter that frequency in the vector element v_j of the vector associated with d_i.

The result is a set of vectors each of which is an occurrence frequency profile of letter tuples for one of the documents in D. These document profile vectors can be stored as the rows of the data matrix.

A letter-pair frequency matrix was abstracted from C using the foregoing procedure. 554 letter pairs were found, and since there are 9 documents, the result is a 9 x 554 matrix henceforth referred to M. An example fragment is shown in Figure 4:

	v1: ic	**v2: ch**	**v3: we**	**...**	**v554: qd**
Sermo Lupi	67	1	86	...	0
Beowulf	400	15	737	...	0
...	...	...	...	...	...
King James Bible	18	18	21	...	0

Figure 4: Letter-pair frequency matrix M abstracted from C

2.2 *Hierarchical cluster analysis of M*

From what is commonly known of the history of the English language and of spelling at various stages of its development, one expects cluster analysis of M to produce no surprises: the Old English, Middle English, and Early Modern English texts will form clusters. This expectation is not fulfilled, however, as is shown in Figure 5:

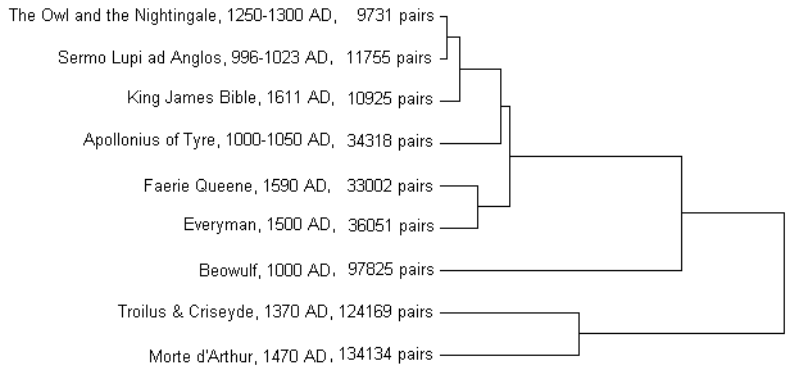


Figure 5: Cluster tree of the rows of data matrix M

The texts do not cluster by chronological period, and the clustering in fact makes no obvious sense in terms of anything one knows about them and their historical context. When, however, one looks at the 'Size' column in Figure 3, the reason for the clustering immediately becomes clear. The texts have been clustered by their relative lengths: the short texts (*Owl, Sermo, King James*) comprise one cluster, the intermediate-length texts (*Apollonius, Faerie Queene, Everyman*) a second cluster, and the long texts (*Troilus, Morte d'Arthur*) a third, with *Beowulf* on its own commensurate with a length that falls between the intermediate-length and long texts.

2.3 *Explanation of document length based clustering*

Clustering based on document length is best explained in terms of vector space geometry, for which see any textbook on linear algebra

such as Fraleigh & Beauregard (1995). A vector space is a geometrical interpretation of a vector in which

- the dimensionality *n* of the vector defines an *n*-dimensional space which, for present purposes, is taken to be the familiar Euclidean one in which the axes are straight lines at right angles to one another.

- the sequence of numbers comprising the vector specifies coordinates in the space.

- the vector itself is a point at the specified coordinates in the space.

For example, the two components of a vector v = (30 70) in Figure 6 are coordinates of a point in a two-dimensional space, and those of v = (40 20 60) of a point in three-dimensional space:

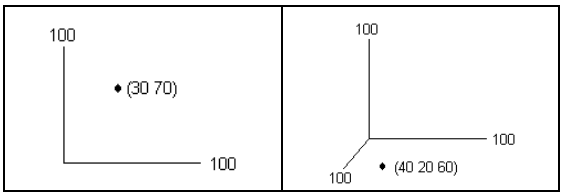


Figure 6: Vectors in two and three dimensional vector spaces

A length-4 vector defines a point in 4-dimensional space, and so on to any dimensionality *n*. Mathematically there is no problem with spaces of dimension greater than 3. The only problem lies in the possibility of visualization and intuitive understanding: as the number of variables and thus dimensions grows beyond 3, graphical representation and intuitive comprehension of it become impossible. The two and three dimensional cases provide a very useful intuitive analogy for higher-dimensional ones, though.

More than one vector can exist in a vector space. Where *n* = 2, for example, a set of vectors in the two-dimensional space might look like Figure 7:

	v1	v2
1	0.8147	0.9058
2	0.1270	0.9134
3	0.6324	0.0975
4	0.2785	0.5469
5	0.9575	0.9649
6	0.1576	0.9706
7	0.9572	0.4854
8	0.8003	0.1419
9	0.4218	0.9157
10	0.7922	0.9595

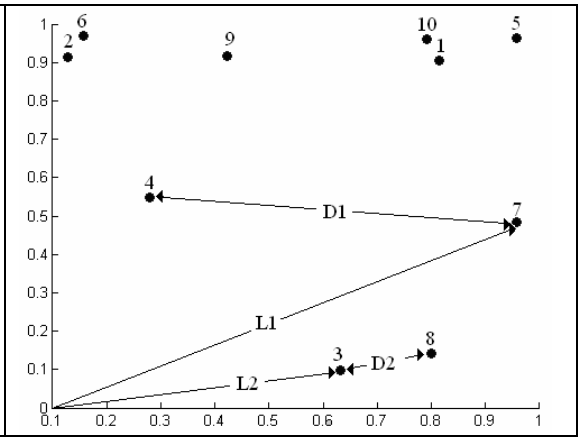


Figure 7: Multiple vectors in two dimensional vector space

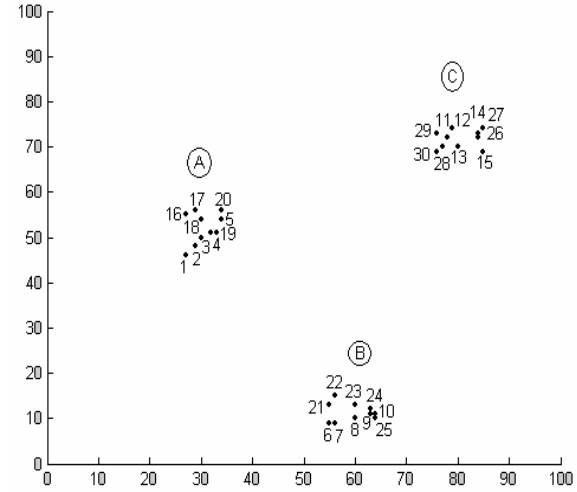
Two concepts associated with vectors in a space are relevant here:

- The length of a vector is the length of a line drawn from the axis origin to the vector's coordinates in the space --for example L1 in Figure 7. Where two or more vectors exist in a space it is possible to compare their lengths: in Figure 7, L1 is greater than L2.

- Where two vectors exist in a space it is possible to measure the distance between them, and when there are more than two their relative distances can be compared; in figure 7 D1, for example, is greater than D2.

Exploratory analytical methods use relativities of vector distance to identify clusters: vectors whose values are relatively similar have similar coordinates in space and are thus relatively close together in the space, whereas vectors whose values are relatively dissimilar are relatively far apart in space. Figure 8 shows a two-dimensional data matrix, the corresponding vectors in two-dimensional space, and a hierarchical analysis showing the cluster structure.

	v1	v2
1	27	46
2	29	48
3	30	50
4	32	51
5	34	54
6	55	9
7	56	9
8	60	10
9	63	11
10	64	11
11	78	72
12	79	74
13	80	70
14	84	73
15	85	69
16	27	55
17	29	56
18	30	54
19	33	51
20	34	56
21	55	13
22	56	15
23	60	13
24	63	12
25	64	10
26	84	72
27	85	74
28	77	70
29	76	73
30	76	69



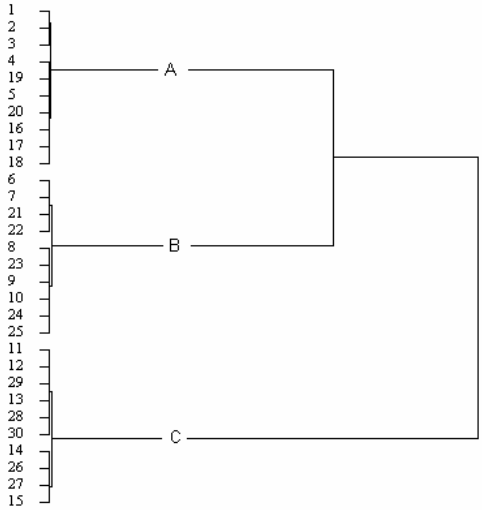
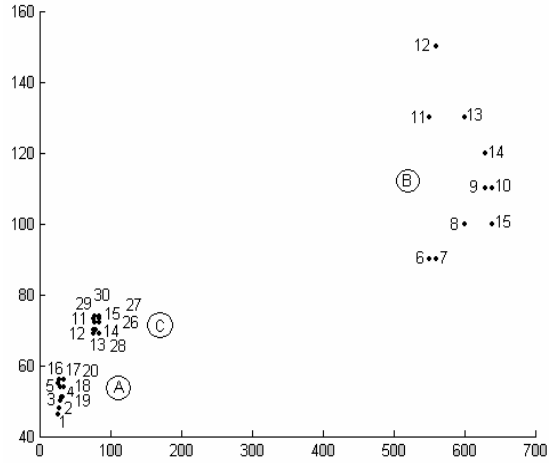


Figure 8: Data matrix with scatter plot of row vectors in two-dimensional space and corresponding cluster tree showing distance relativities of row vectors in the space

Now observe what happens to the distribution of the row vectors in the space and the corresponding cluster tree when a proper subset of them is lengthened. A vector is lengthened by increasing the magnitude of the numbers that comprise it. The numerical values of all the vectors belonging to cluster B in Figure 8 were multiplied by 10; note that this is a random selection both of vectors and of multiplier, and the discussion to follow would have been the same with a different selection. The resulting matrix, together with the corresponding scatter plot and cluster tree, are shown in Figure 9:

	v1	v 2
1	27	46
2	29	48
3	30	50
4	32	51
5	34	54
6	550	90
7	560	90
8	600	100
9	630	110
10	640	110
11	78	72
12	79	74
13	80	70
14	84	73
15	85	69
16	27	55
17	29	56
18	30	54
19	32	51
20	34	56
21	550	130
22	560	150
23	600	130
24	630	120
25	640	100
26	84	72
27	85	74



28	77	70
29	76	73
30	76	69

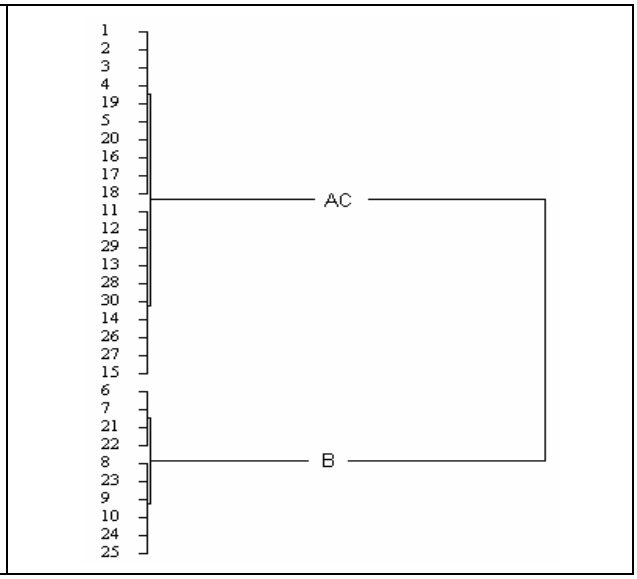


Figure 9: Modified matrix from Figure 8 with corresponding scatter plot and cluster tree

Lengthening the row vectors of cluster B in Figure 8 has moved them far from A and C and brought A and C relatively much closer together. The consequence for clustering is shown in the corresponding tree, which now differs fundamentally from the one in Figure 8: A and C now form a composite cluster, and B is far from AC. In this case, therefore, it is clear that relative vector length is an important determinant of clustering, and, more specifically, that vectors of similar lengths cluster --long with long and short with short. The general case is not quite so simple, since the angles between and among vectors and not just their relative lengths also need to be taken into account (Fraleigh & Beauregard 1995), and it is more accurate to say that, in general, vectors of similar lengths *tend* to cluster.

How does all this apply to length-based clustering of varying-length document collections? When, as here, the data abstracted from a collection is a frequency matrix based on counting all occurrences of a set of features in each text, the sum of magnitudes of the frequencies in the vector representing a long document will be greater than the

sum of magnitudes of frequencies in the vector representing a short one --or, on other words, vectors representing long documents are longer than vectors representing short documents in a way that is proportional to the difference in document lengths. This is shown in Figure 10, where row vector lengths in M are plotted against the lengths of the corresponding documents in C, and where vector length grows near-linearly with document length:

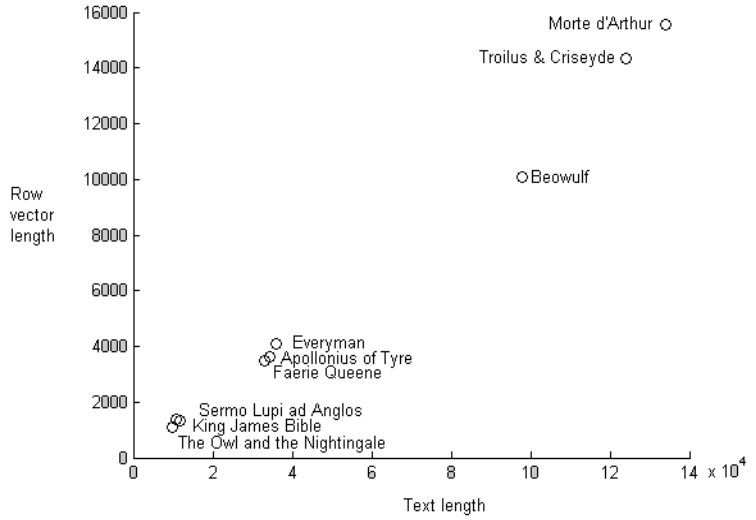


Figure 10: Plot of row vector lengths in M against the lengths of the corresponding documents in C

Comparison of Figure 10 with the cluster tree in Figure 5, moreover, shows an isomorphism between the vector length relations in Figure 10 and the document clustering in Figure 5: the documents have been clustered by relative vector length.

3. Solutions

There is an obvious solution to the problem of variation in document length: truncate all the documents to the length of the shortest, thereby

making them all equal in length. There are, however, two problems with this approach. On the one hand, where the variation is large and the shortest documents are very short, it entails throwing away a good deal of potentially useful information. And, on the other, there is no obvious basis for choosing what material to retain from the longer texts and what to discard. For these reasons, alternatives to truncation have been developed.

The literature contains a variety of ways of mitigating or eliminating the effect of variation in document length on data matrix row vector clustering (Buckley 1993; Singhal *et al*. 1996a, 1996b). We will consider the one that is probably the intuitively most accessible: normalization by mean document length. This normalization adjusts the lengths of each row vector of an *m* x *n* frequency matrix, here M, in relation to the mean length of documents in the collection:

$$M'_i = M_i \times \frac{\mu}{length(i)}$$

where:

- M'_i is the normalized *i*'th row vector of the matrix M, for *i* = 1..the number of rows *m* in M.
- M_i is the unnormalized *i*'th row vector of the matrix M
- μ is the mean number of letter pairs across the *m* documents
- *length*(*i*) is the number of letter pairs in any given document *i*

The value in each document vector M_i is multiplied by the ratio of the mean number of letter pairs across all the documents in the collection to the number of pairs in document *i*. The effect is to decrease the values in the vectors that represent long documents, increase them in vectors that represent short ones, and, for documents that are near or at the mean, to change the corresponding vectors little or not at all. Conceptually, therefore, this normalization constitutes a conjecture about what the row vectors in a data matrix would have been like if the corresponding documents had all been the same length.

Cluster analysis of the normalized matrix M' is shown in Figure 11:

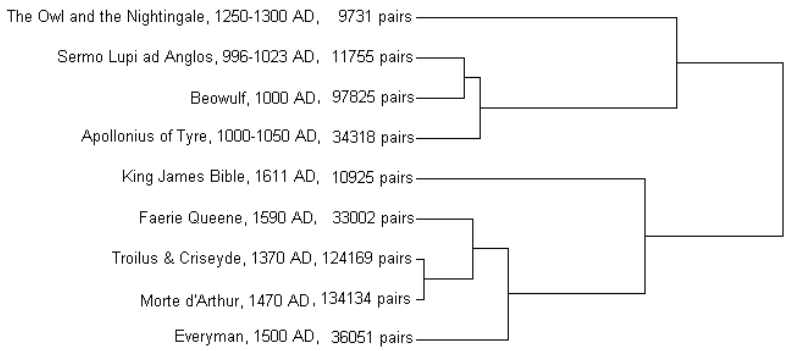


Figure 11: Cluster tree of the rows of length-normalized matrix M'

The row vectors are now clustered by the chronological periods of the texts they represent, and make sense in terms of what is known of those texts in relation to the history of English. There are two main clusters. The upper one subclusters into a group of Old English texts and the single Early Middle English text irrespective of length variation. The lower one contains the later Middle English and the Early Modern English texts. Here, the most recent of the Early Modern texts, *King James*, is on its own; the *Faerie Queene*, though chonologically near to *King James*, is known deliberately to have archaized its spelling, and is thus classified with the Middle English texts. Document length normalization has, therefore, solved the problem of clustering by document length in this instance. The NECTE data matrix discussed in Section 1 above was, moreover, normalized prior to cluster analysis, and the tree shown in Figure 2 is based on the normalized matrix.

4. Discussion

Document length normalization is not as straightforward as the foregoing discussion suggests, for two main reasons.

Firstly, the normalization procedure used in Section 3 solved the problem of variation in document length in the sense that, for the small example corpus C, it supported a cluster analysis that gave the expected answer, and, for NECTE, supported an analysis that is sociolinguistically plausible. But how does its performance compare to the other available normalization methods, both with respect to these and to more general applications? Selection of an appropriate method must be based on an evaluation of their relative effectiveness; the plan is to undertake such an evaluation as part of future research on document length normalization.

Secondly, frequency matrices based on collections of varying-length documents can have characteristics that compromise the effectiveness of existing normalization procedures. One of these is nonlinearity in the growth of variable frequency with increasing text length, and another is unreliable population probability estimation for variables in very short documents. An adequate account of the former would excessively prolong the discussion and is therefore not attempted here, but see Moisl (2007) for an indication of what is involved. A brief account of the latter follows.

Given a population E of n events, the frequency interpretation of probability (Milton & Arnold 2003:1-17) says that the probability $p(e_i)$ of $e_i \varepsilon$ E (for i in 1..n) is the ratio (*frequency* (e_i) / n), that is, the proportion of the number of times e_i occurs relative to the total number of occurrences of events in E. For example, if a document contains 100,000 letters and the letter g occurs 320 times, then the probability $p(g)$ = 320 / 100000. A sample of E can be used to estimate $p(e_i)$, as is done with, for example, human populations in social surveys. The Law of Large Numbers (Grinstead & Snell 1997: 305-320) says that, as sample size increases, so does the likelihood that the sample estimate of an event's population probability is accurate; a small sample might give an accurate estimate but is less likely to do so than a larger one, and for this reason larger samples are preferred.

With specific reference to document corpora, it was pointed out earlier that, where the data abstracted from a multi-document corpus is a frequency matrix based on counting all occurrences of a set of features in each document, the sum of magnitudes of the frequencies in a vector representing a relatively longer document is

greater than the sum for the vector representing a relatively shorter one. The longer the document, therefore, the more accurate its estimation of the population probabilities of the selected textual features can be expected to be. To exemplify this, a randomly selected document --Dickens' *Domby & Son*-- was partitioned into a corpus D of 100 increasing-length segments: the first segment contains the first 1000 words of the novel, the second segment the first 2000 words, and so on, adding the next 1000 words to segment *i* to create segment *i* + 1. A matrix Q of letter-pair frequencies was abstracted from D as in Section 2.1 above. For convenience of exposition, the matrix rows were arranged in ascending order of row vector length so that the one representing the shortest segment was at Q_1 and the longest at Q_{100}, and the columns so that the highest-frequency variable was represented by the leftmost column and the lowest frequency variable in the rightmost one. The probabilities for each of the letter-pair columns of Q were then calculated to find out the relationship between segment length and accuracy of population probability estimation for each pair across the entire 100-segment collection. The probability distributions for the three most frequent pairs *he*, *th*, and *in* are shown in Figure 12; the distributions for the remaining columns are similar.

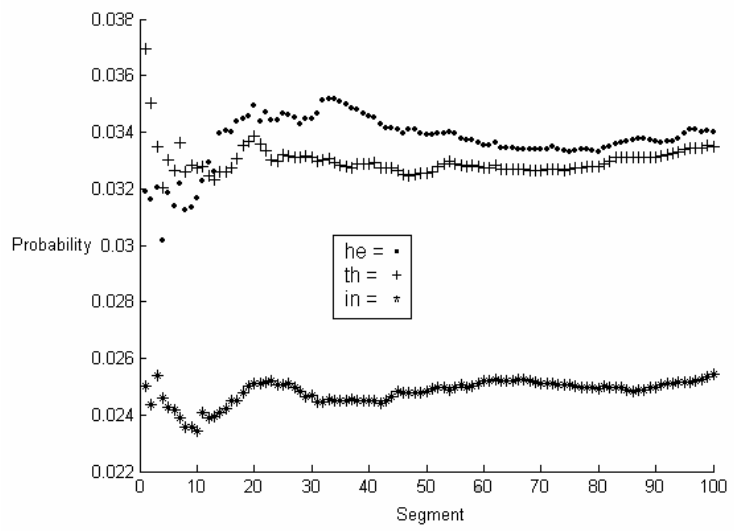


Figure 12: Probability estimates of the letter pairs *he*, *th*, and *in*, where the horizontal axis represents segments of increasing length and the vertical axis probability.

The horizontal axis represents the 100 segments and the vertical axis the probability estimates for *he*, *th*, and *in*. In each distribution, the probabilities fluctuate for the shorter segments on the left and then settle down to a fairly constant value representing the increasingly-accurate estimate of the population probability as one moves to the longer segments on the right, which is what one expects from the Law of Large Numbers. The fluctuations on the left are caused by frequency values that are too large or too small relative to the length of the segment to estimate the population probability accurately. In other words, frequency values for variables in short texts can be and in the present instance are unreliable estimators of population probabilities.

This unreliability can render document length normalization unreliable as well. To show how, Q was normalized using the same procedure as in Section 3, and the effect on the values in the *he* column is shown in Figure 13.

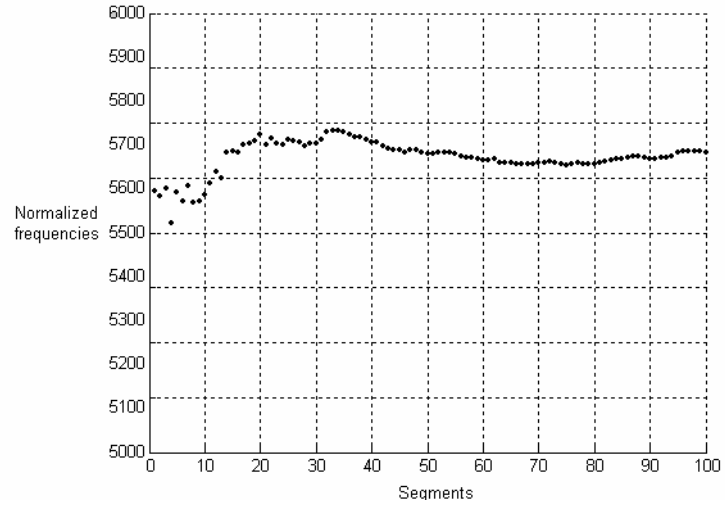


Figure 13: Normalized values for the latter pair *he*, where the x-axis represents texts of increasing length and the y-axis the normalized values.

Figure 13 shows a normalized frequency distribution curve isomorphic with the probability curve in Figure 12: for the shorter segments, the normalized values fluctuate between about 5100 and 5900 before settling down to a value around 5600. This degree of variation can be expected to affect assignment of the shorter segments to clusters in cluster analysis. The suspicion, moreover, is that Q is not unique in this respect, and that the effect just described will occur for matrices derived from other document collections --that, in short, this is a general problem in document length normalization. How widespread it is, and how important its effect on exploratory analysis, is a matter for further research. And, if it is both widespread and important, so is what to do about it.

5. Conclusion

To use exploratory multivariate methods effectively in the analysis of document collections, issues that arise with respect to the abstraction of data from such collections have to be understood. This paper addressed an issue that has a fundamental bearing on the validity of analytical results based on such data: variation in the lengths of the documents in the collection of interest. The discussion was in four main parts. The first part showed how a particular class of computational methods, exploratory multivariate analysis, can be used in historical dialectology research, the second explained why variation in document length can be a problem in such analysis, the third presented a solution --normalization of document length relative to the mean length of documents in the collection-- and the fourth pointed out some difficulties that arise in relation to document length normalization. The conclusion is that failure to normalize for variation in document length can generate fundamentally erroneous cluster analytical results, but that normalization itself has some unresolved problems.

5. References

Allen W. / Beal J. / Corrigan K. / Maguire W. / Moisl, H. 2006. A linguistic "time capsule": the Newcastle Electronic Corpus of Tyneside English. In Allen W. / Beal J. / Corrigan K. / Maguire W. / Moisl, (eds) *Creating and Digitizing Language Corpora, Volume 2: Diachronic Databases*. Basingstoke, UK: Palgrave Macmillan, 16-48.

Andrienko, N. / Andrienko, G. 2005. *Exploratory Analysis of Spatial and Temporal Data: A Systematic Approach*. Heidelberg: Springer.

Belew, R. 2000. *Finding out about: A cognitive perspective in search engine technology and the WWW*. Cambridge: Cambridge University Press.

Bishop, C. 2006 *Pattern Recognition and Machine Learning*. New York: Springer.

Buckley, C. 1993. The importance of proper weighting methods. In Bates, M. (ed.) *Human Language Technology.* San Mateo, CA: Morgan Kaufmann.

Everitt, B. / Landau, S. / Leese, M. 2001. *Cluster Analysis*, 4th ed. London: Arnold.

Fraleigh, J. / Beauregard, R. 1995. *Linear Algebra.* 2nd ed. Menlo Park, CA: Addison-Wesley.

Grinstead, C. / Snell, J. 1997. *Introduction to Probability*, 2nd ed. American Mathematical Society.

Grossman, D. / Frieder, O. 2004. *Information Retrieval.* 2nd ed. Dordrecht: Springer.

Hair, J. / Black, W. / Babin, B. / Anderson, R. / Tatham, R. 2005. *Multivariate Data Analysis.* 6th ed. New Jersey: Prentice-Hall.

Heeringa, W. / Nerbonne, J. 2001. Dialect areas and dialect continua. *Language Variation and Change* 13, 375-400.

Milton, J. / Arnold, J. 2003. *Introduction to Probability and Statistics*, 4th ed. Boston:McGraw-Hill.

Moisl, H. / Jones V. 2005. Cluster analysis of the Newcastle Electronic Corpus of Tyneside English: a comparison of methods. *Literary and Linguistic Computing* 20, 125-46.

Moisl, H. / Maguire W. / Allen W. 2006. Phonetic variation in Tyneside: exploratory multivariate analysis of the Newcastle

Electronic Corpus of Tyneside English. In Hinskens, F. (ed.) *Language Variation. European Perspectives*. Amsterdam: John Benjamins, 127-141.

Moisl, H. 2007. Data nonlinearity in exploratory multivariate analysis of language corpora,. In Nerbonne, J. / Ellison, M. / Kondrak, G. (eds) *Computing and Historical Phonology. Proceedings of the Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology*, June 28 2007, Association for Computational Linguistics, 93-100.

Moisl, H. / Maguire, W. 2008. Identifying the Main Determinants of Phonetic Variation in the Newcastle Electronic Corpus of Tyneside English. *Journal of Quantitative Linguistics* 15, in press.

Nerbonne J. / Heeringa W. 2001. Computational comparison and classification of dialects. *Dialectologia et Geolinguistica* 9, 69-83.

Singhal, A. / Salton, G. / Mitra, M. / Buckley, C. 1996a. Document Length Normalization. *Information Processing and Management* 32, 619-633.

Singhal, A. / Buckley, C. / Mitra, M. 1996b. Pivoted document length normalization. *Proceedings of the 19th ACM Conference on Research and Development in Information Retrieval (SIGIR-96)*, 21-29.

Tabachnik, B. / Fidell, L. 2006. *Using Multivariate Statistics*, 5th ed.. Boston: Allyn & Bacon.

Tan, P. / Steinbach, M. / Kumar, V. 2006. *Introduction to Data Mining*. Boston: Pearson Addison Wesley.