

Data Normalization for Variation in Document Length in Exploratory Multivariate Analysis of Text Corpora

Hermann Moisl
School of English and Linguistics
University of Newcastle
Newcastle upon Tyne NE1 7RU, UK
Hermann.Moisl@ncl.ac.uk

Abstract

The advent of large electronic text corpora has generated a range of technologies for their search and interpretation. Variation in document length can be a problem for these technologies, and several normalization methods for mitigating its effects have been proposed. This paper assesses the effectiveness of such methods in specific relation to exploratory multivariate analysis. The discussion is in four main parts. The first part states the problem, the second describes some normalization methods, the third identifies poor estimation of the population probability of variables as a factor that compromises the effectiveness of the normalization methods for very short documents, and the fourth proposes elimination of data matrix rows representing documents which are too short to be reliably normalized and suggests ways of identifying the relevant documents.

1. Introduction

The advent of large electronic text corpora has generated a range of technologies for their search and interpretation. Variation in document length can be a problem for these technologies, and several normalization methods for mitigating its effects have been proposed. This paper assesses the effectiveness of such methods in specific relation to exploratory multivariate analysis [8, 15]. The discussion is in four main parts. The first part states the problem, the second describes some normalization methods, the third identifies poor estimation of the population probability of variables as a factor that compromises the effectiveness of the normalization methods for very short documents, and the fourth proposes elimination of data matrix rows representing documents which are too short to be reliably normalized and suggests ways of identifying the relevant documents

2. Variation in document length: the problem

Documents in collections can and often do vary considerably in length. Where the data abstracted from such a collection is based on the frequency of some textual feature or features of interest, such length variation is a problem for exploratory multivariate analysis. The nature of the problem is exemplified using the small document collection C comprising excerpts of various lengths from historical English texts ranging from Old English to Early Modern English, shown in Table 1.

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

Table 1. Document collection C

2.1 Data creation

Prior to its standardization in the later 18th century, spelling in the British Isles varied considerably over time and 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. The research question, therefore, is: can the documents in C be accurately categorized chronologically by their spelling?

Investigation of spelling is here based on the concept of the tuple, where a tuple is a sequence of symbols: xx a pair, xxx a triple, $xxxx$ a four-tuple, and so on. Given a collection containing m documents, compile a list of all

letter tuples that occur in the texts. Assume that there are n such tuples. To each of the documents d_i in the collection (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 the collection. These document profile vectors are stored as the rows of a 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 9×554 matrix henceforth referred to M_C .

2.2 Exploratory multivariate analysis of M_C

From what is commonly known of the history of the English language and of spelling at various stages of its development, one expects exploratory analysis of M_C to produce no surprises: the Old English, Middle English, and Early Modern English texts will form clusters. This expectation is not fulfilled, however, as the hierarchical analysis [5] in Figure 1 shows.

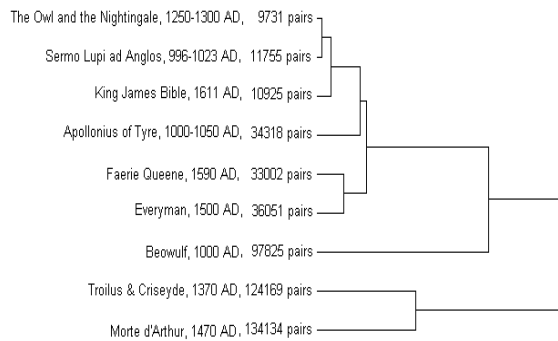


Figure 1. Cluster tree of the rows of data matrix M_C

The texts do not group 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 Table 1, a correlation between cluster structure and document length is immediately clear. The texts have been grouped 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

When data has a vector representation, clustering by document length is explicable in terms of vector space geometry [6], in which the dimensionality n of the vector defines an n -dimensional space (here taken to be the familiar Euclidean one), the sequence of scalars comprising the vector specifies coordinates in the space, and the vector itself is a point at those coordinates. When two or more vectors exist in a space it is possible to measure the distance between them and thus to compare relative distances, so that $distance(AB)$ in Figure 2a is greater than $distance(AC)$. The length of a vector is the distance between itself and some reference point in the space's coordinate system; for present purposes that reference point is taken to be the origin of the coordinate axes. Like the distance between vectors, the relative lengths of vectors can be compared --in Figure 2b $length(A)$ is greater than $length(C)$.

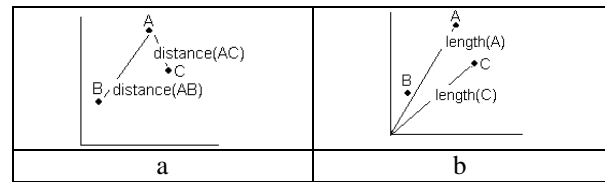


Figure 2: Distance and length in two-dimensional vector space

The distance between any two vectors in a space is jointly determined by the magnitude of the angle between the lines joining them to the origin of the space's coordinate system, and by the lengths of those lines.

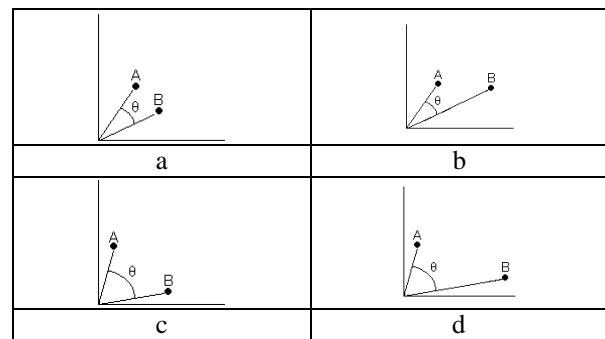


Figure 3: Relationship of vector angle and vector length to vector distance

Figure 3a shows two vectors A and B and an angle θ between them. In 3b θ remains the same and the length of B is increased, in 3c θ is increased and the vector lengths remain the same, and in 3d both the angle and the length

of B are increased; in all cases (3b) - (3d) the distance between A and B increases commensurately.

It is easy to see that, as the angle decreases, length becomes increasingly dominant in determining distance. When, moreover, this observation is extended to more than two vectors, length becomes an increasingly important determinant of vector clustering in the space: where the angles between them are small, vectors of similar lengths cluster, as shown in Figure 4.

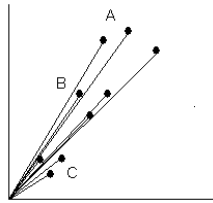


Figure 4: Clusters determined by vector length

And, because hierarchical cluster analysis groups vectors on the basis of their relative distances in space, vector length under these circumstances largely determines cluster analytical results.

This applies directly to the cluster analysis of M_C in that (i) the angles between its row vectors are relatively small, (ii) the vectors vary in length, and (iii) this length variation creates clusters in the data space. Because M_C is 554-dimensional there is no question of being to show this by plotting the row vectors directly as for the two-dimensional example in Figure 4. It is, however, possible to do so indirectly by projecting M_C into two-dimensional space using principal component analysis [9] and then plotting the rows of the projection matrix; the two largest principal components of M_C account for 70.7% of its variance, so the 9×2 projection matrix $M_{C(PCA)}$ is a reasonably accurate representation of M_C . The scatter plot of the rows of $M_{C(PCA)}$ in Figure 5 shows that the angles between them are indeed relatively small and that they cluster by vector length.

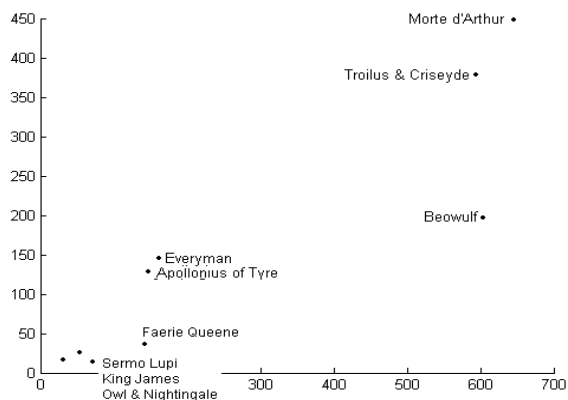


Figure 5: Scatter plot of the row vectors of $M_{C(PCA)}$

When, moreover, one observes that there is a near-linear relationship between the sizes of the documents in C (measured as the number of tuples in each) and the lengths of the vectors representing them in M_C (Figure 6), the reason for the length-based clustering of the documents in C becomes obvious.

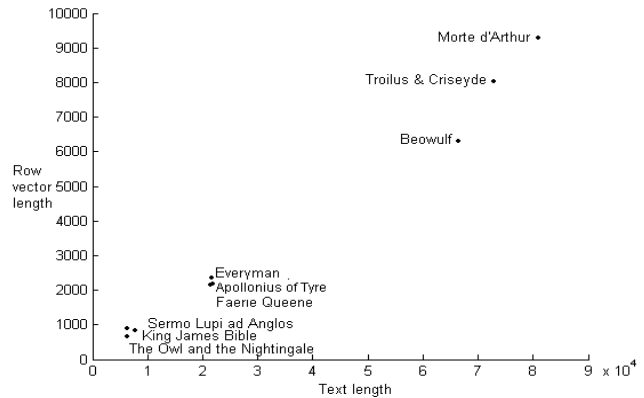


Figure 6: Plot of row vector lengths in M_C against the sizes of the corresponding documents in C

3. Document length normalization methods

Several ways of normalizing frequency data matrices abstracted from varying-length document collections have been proposed [2, 13, 14]. All of them work by dividing each of the n values in each of the m rows of a frequency matrix M by a constant: k :

$$M_{ij} = \frac{M_{ij}}{k}$$

for $i = 1..m, j = 1..n$. This section mentions only two; subsequent discussion will show why an exhaustive list is unnecessary for present purposes.

- *Probability normalization*: For a given row M_i , k is the sum of frequencies in that row, that is, $k = \sum_{j=1..n} M_{ij}$. This replaces absolute frequency values in the matrix, whose magnitudes are dependent on document size, with probabilities, which are not; see further on the frequency-based definition of probability in Section 4 below.
- *Cosine normalization*: Any vector can be transformed so that it has length 1 by dividing it by its norm or length:

$$v_{unit} = \frac{v}{|v|}$$

In the present application $M_i = v$ and $|M_i| = k$. All row vectors in M are thereby made to lie on a hypersphere of radius 1 around the origin; because all vectors are equal in length, variation in the lengths of documents and, correspondingly, of the vectors that represent them cannot be a factor in analysis.

4. Effectiveness of normalization methods

M_C was normalized using the methods described in Section 3, and both the normalized matrices were cluster analyzed. In both cases the result was the same, and is shown in Figure 7.

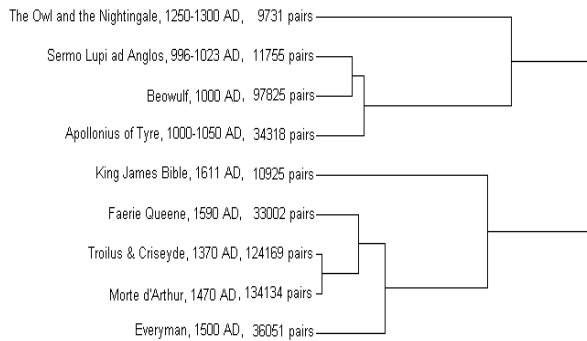


Figure 7. Cluster analysis of length-normalized matrix M_C

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 comprises 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 chronologically near to *King James*, is known deliberately to have archaized its spelling, and is thus classified with the Middle English texts.

For C , therefore, the conclusions are (i) that normalization solves the problem of variation in document length, and (ii) that the normalization methods referred to in Section 3 are equally effective. Can these conclusions be extended to document collections in general? The short answer with respect to (i) is 'no', and with respect to (ii) 'probably'; the remainder of this section deals mainly with (i), but (ii) is briefly addressed at the end.

When a frequency matrix is abstracted from a collection containing very short documents, normalization

of the vectors representing those short documents is likely to be unreliable, which in turn leads to unreliable cluster analytical results. This stems from the unlikelihood of very short texts accurately estimating the population probabilities of data variables. Given a population E of n events, the frequency interpretation of probability [11, pp.1-17] says that the probability $p(e_i)$ of $e_i \in E$ (for $i = 1..n$) 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 . 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 [7, pp. 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. It has already been pointed out that, where there is variation in document length and all occurrences of some feature are counted, the sum of frequencies for a vector representing a relatively longer document is necessarily greater in magnitude than the sum of frequencies for a vector representing a relatively shorter one. The shorter the document, therefore, the less accurate its estimate of the population probabilities can be expected to be.

To see the effect of this on cluster analysis, consider first a case where the population probabilities of the data variables are known, and a data matrix where the rows represent samples s of increasing size and the sample variable values have been arranged so that all give perfect estimates of those probabilities (Table 2).

	v1 p = .067	v2 p = .133	v3 p = .200	v4 p = .267	v5 p = .333
r1 (s=15)	1	2	3	4	5
r2 (s=30)	2	4	6	8	10
r3 (s=60)	4	8	12	16	20
r4 (s=120)	8	16	24	32	40
r5 (s=240)	16	32	48	64	80
r6 (s=480)	32	64	96	128	160
r7 (s=960)	64	128	192	256	320
r8 (s=1920)	128	256	384	512	640

Table 2. Matrix showing population probabilities of variables

Table 3 shows this matrix probability-normalized.

	v1	v2	v3	v4	v5
r1	.067	.133	.200	.267	.333
r2	.067	.133	.200	.267	.333
r3	.067	.133	.200	.267	.333
r4	.067	.133	.200	.267	.333
r5	.067	.133	.200	.267	.333
r6	.067	.133	.200	.267	.333

r7	.067	.133	.200	.267	.333
r8	.067	.133	.200	.267	.333

Table 3. The matrix of Table 2 probability-normalized

The matrices of Tables 2 and 3 were cluster analyzed, and the results are shown in Figure 8.

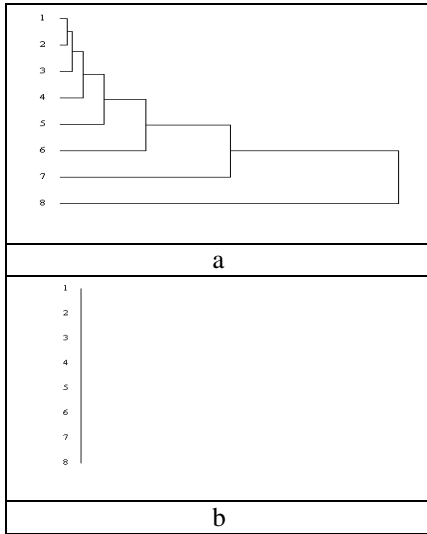


Figure 8. Cluster analyses of Figures 10 (a) and Figure 11 (b) matrices

Normalization has completely eliminated the variation in length which gives rise to the length-based clustering in Figure 8a and made the rows unclassifiable (8b), as the definition of probability normalization leads one to expect.

Now consider what happens with a matrix empirically derived from a collection of, say, 16 documents where accuracy of the population probability estimates cannot be guaranteed. For comparability with Tables 2 and 3, each document in the collection is twice as long as the preceding one, giving the same progression of relative sample lengths as in Table 2. The documents are increasing-length excerpts from a randomly-selected text, Dickens' *Dombey & Son* [12], and the variables are again letter pairs: the first document contains the first 10 letter pairs in the text, the second the first 20 pairs, and so on up to the sixteenth at 327680 pairs. There are 560 letter-pair types, which yields a 16 x 560 frequency matrix M_{560} . Figure 9a shows a cluster analysis of M_{560} , and Figure 9b of the probability normalized matrix $M_{560(norm)}$.

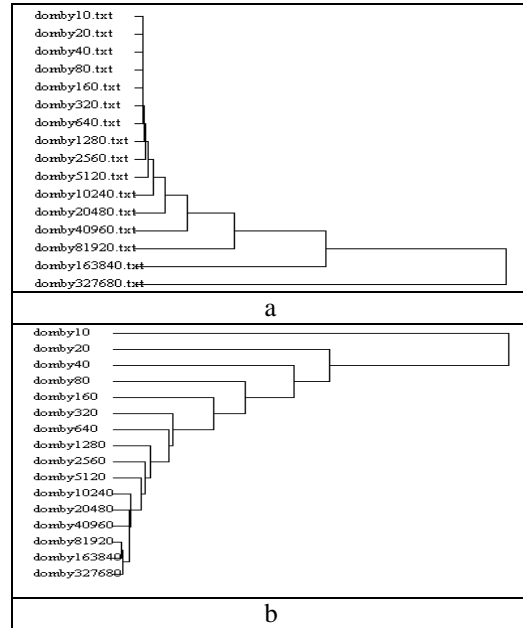


Figure 9. Euclidean distance / single link cluster analysis of M_{560} and $M_{560(norm)}$

Like Figure 8a, 9a shows length-based clustering. Unlike Figure 8b, however, 9b is not flat, that is, the matrix rows have not been normalized to uniform values. The reason for this emerges from an examination of the distributions of individual variable probabilities. Figure 10 shows the distributions for the three most frequent letter pairs in the collection, *th*, *in*, and *he*, across all 16 documents; the remaining columns are similar.

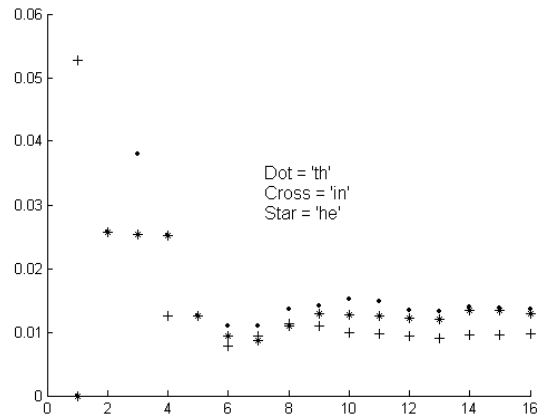


Figure 10. Probability distributions of the letter pairs *he*, *in*, and *th*

The horizontal axis represents the 16 documents with the shortest on the left and the vertical axis the population probability estimates for *he*, *th*, and *in*. In each distribution, the probabilities fluctuate for the shorter

documents and then settle down to a much more restricted range of values corresponding to the increasingly-accurate estimate of the population probability as one moves to the longer documents 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 text sample 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.

Finally, it remains to note that this unreliability of normalization with respect to very short documents affects any method that divides row vector values by a constant, such as the cosine normalization mentioned in Section 3. These methods are all linear vector transformations, and, as such, affect the scaling of the row values but not their distribution.

5. Dealing with very short documents

The obvious solution to the problem described in the preceding section is to determine which documents in a collection are too short to provide reasonably reliable estimates of population probabilities, and to eliminate the corresponding rows from the data matrix. But how short is too short? The answer proposed in this section is implicit in Figure 10: because documents that are too short to give reliable probability estimates have large vertical fluctuations, the point on the horizontal axis where the fluctuations settle down is the required document length threshold. In Figure 10, that means documents 1-3. One would, of course, want to look at more variables to confirm this, but the principle remains the same.

The collection on which Figure 10 is based was, however, selected to make the point made in section 4 as clearly as possible, and is unrepresentative of corpora likely to be analyzed using exploratory multivariate methods in actual research applications. In particular, the 16 'documents' are in fact samples taken from the same single-authored text, and, assuming that that author's spelling, prose style, and subject matter are broadly constant across the text as a whole, the rapid convergence on specific population probabilities for each of the variables shown in Figure 10 is unsurprising. But in many and probably most actual research applications the corpora being analyzed are more disparate --they may differ in some combination of such things as genre, subject matter, date, and authorship-- and as such the same clear convergence on the population probabilities of the textual features of interest will not necessarily obtain. Will the approach to document length threshold determination just proposed work equally well in such cases?

To test this, a relatively large and disparate corpus was assembled. It consists of 134 Middle English documents

selected at random from those available online at the *Corpus of Middle English Prose and Verse* website [4], and is intended to be more representative of the kind of collection that an historian or an historical linguist might be interested in analyzing. The documents range in date from about 1250 to about 1500 CE, were written for the most part by different authors --some known and some unknown-- in various dialect regions of Britain, and vary in length from 1Kb to 1420 Kb. A matrix of letter pair frequencies M was constructed in which each row M_i represents a different document, each column M_j a different letter pair, and each M_{ij} the number of times letter pair j occurs in document i ; 883 letter pairs were found, yielding a 134 x 883 matrix. To facilitate the discussion that follows the rows were sorted in increasing order of length so that the one representing the shortest document was at $M_{1,j}$, and the columns were sorted in decreasing order of summed frequency so that the one representing the most frequent letter pair was at $M_{i,1}$. The sorted matrix was then probability-normalized and the values in the columns for each of the 134 documents were plotted, as for Figure 10, so that the normalized probability values are on the vertical axis and the document numbers on the horizontal one. Figure 11 shows the plot for *er*, the fourth most frequent letter pair in the corpus; the plots for other frequent pairs are similar.

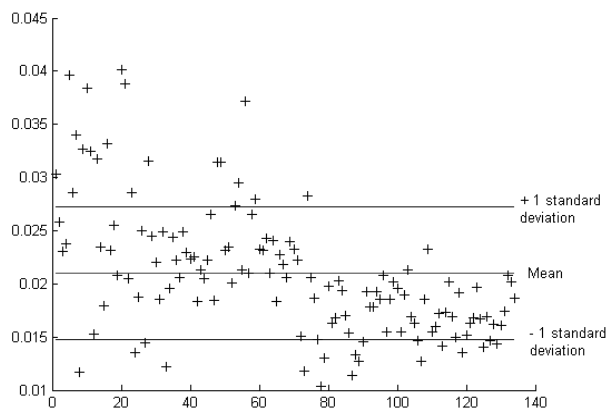


Figure 11: Probability normalized values for letter pair *er* across all corpus documents

The middle line shows the mean of the normalized values on the vertical axis across all 134 documents and the upper and lower lines the standard deviation of those values; these come into the discussion shortly. The first thing to note is that the probability estimates for *er* do not converge as quickly or as neatly as for the variables in Figure 10: there is substantial variation in the estimates even for the longest documents, and this presumably reflects the disparate character of the collection. A convergence from the shorter documents on the left of the

plot to the longer ones on the right is nevertheless clearly visible, and to that extent the approach to threshold determination proposed with respect to Figure 10 works here as well. But where to place the threshold is by no means as obvious as in Figure 10. Working purely from visual intuition, one could justify any threshold between about document 25 and document 80; with such a large range one can easily eliminate or retain too many documents, to the detriment of the analysis. What is required is a more principled guide to threshold selection.

Statistical sampling theory provides such a guide. A fundamental question in sampling theory is: 'How large does a sample have to be to estimate the value of a parameter μ for some population characteristic of interest with reasonable accuracy?'. [6] provides an overview of approaches to this question, [3] is a classic work on the subject, and [1, 5, 12] are more or less recent and accessible discussions. In cases where the parameter to be estimated is the proportion of the characteristic in the population, the answer is:

$$n = \frac{t^2(p)(1-p)}{e^2} \quad (1)$$

where

- n is the size of the required sample.
- p is an estimate of the population proportion μ .
- e is the error tolerance $\mu - p$ that one is prepared to accept in a sample estimate of the population value of μ .
- t is the 'alpha' or significance level which specifies the probability that an estimate p is within a tolerance e of μ . This t is not stated directly as a probability, that is, as a real number in the range 0..1, but as the number of standard deviations from μ within which p will fall with the specified probability. A frequently used value for t is 1.96, which corresponds to a 95% chance that p will fall within the tolerance e . The validity of the t element in expression (1) depends on the assumption of normality in the distribution of values in the sample.

Applying these ideas to the identification of a document length threshold, the population from the point of view of, say, an historical linguist wishing to use the 134 documents in the corpus as a basis for generalization about Middle English spelling is the set of all English-language documents written in Britain between about 1100 and 1500 CE, the population characteristic of interest is the letter pair er , the parameter being estimated is the probability or, equivalently, the proportion of er in the population, and the problem is to find a document length n , expressed as the number of letter pairs it contains, which will provide a reliable estimate of the population

probability of er . This problem is addressed by regarding each of the 134 documents as a sample from the population, and each column of the data matrix M as a sampling distribution of the probability of the associated letter pair. The Central Limit Theorem says that, as the number of samples grows, the sampling distribution approaches normality and the estimates of population mean and standard deviation become increasingly accurate. The mean, standard deviation, and shape of the sampling distribution for er are shown in Figure 12:

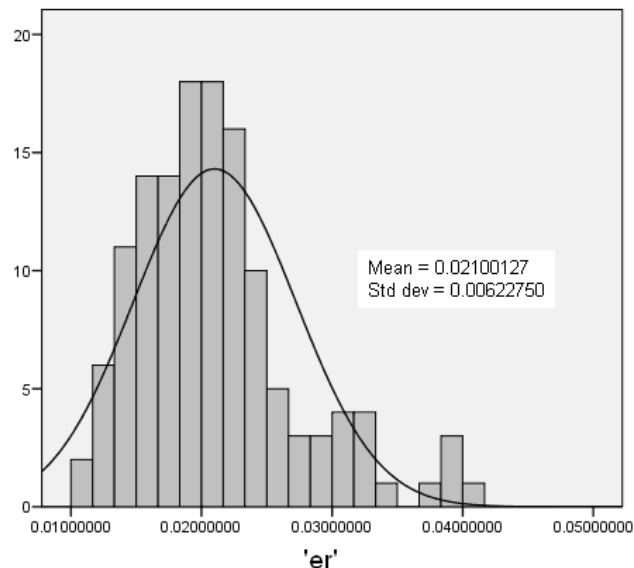


Figure 12: Sampling distribution for er

The distribution is approximately normal and so reasonably well satisfies the normality assumption which underlies the t component of expression (1); the conventional value $t = 1.96$ is used in what follows. Given the fairly large number of samples, the mean of the sample probabilities is taken to be a good estimate of the population probability for er , and is used as the value for p in expression (1). The acceptable error e is analyst-defined, so the question of what a suitable value might be arises in the present case. Figure 11 shows a convergence on values that fall within 1 standard deviation of the mean, and that is taken to be an acceptable value for e . Then:

$$n = \frac{(1.96)^2(0.00100127)(1-0.00100127)}{(0.00622750)^2} \approx 2052$$

For $t = 2.58$ corresponding to a 99% confidence level, $n = 3555$. Working from the shortest of the 134 documents to the longest, it turns out that document 25 contains 1976 letter pairs and the next longest one 26 contains 2240: the lower bound for the required length threshold is document

25, and the upper bound, corresponding to $n = 3555$, is document 56. These bounds accord well with visual intuitions about Figure 11, in that the lower one eliminates the documents with very large fluctuations on the very left of the plot, and the upper one eliminates most of those with the medium-level fluctuations.

Clearly, a decision on document length threshold based only on one letter pair variable is unreliable, and the foregoing procedure would have to be applied to a reasonable number of other frequent variables to arrive at a consensus, as recommended by [3, ch. 4].

Finally, two comments. The first is that arguments based on sampling distributions assume that the samples are of equal size. This is manifestly not the case for our corpus, and would appear to undermine what has been said above. The problem is only apparent, however, because the matrix on which the calculations are based was length-normalized. Secondly, the sampling distribution for er in Figure 12 is skewed, and skewness affects the accuracy of sample size estimation [3, pp.39-44]. Its cause can be seen in Figure 11, where normalization of the poor probability estimates by the shortest documents has generated values that are much larger than those on which er converges. It is clear from Figure 11 that the convergence is to the bottom half of the e range, or, in other words, that the mean is too high --it should be about halfway between the current mean and lower standard deviation lines in Figure 11, and the standard deviation used for e in the calculation of n should be relative to that. An actual corpus analysis would need to recalculate n taking the skewness into account, but this is a methodological paper, and it was felt to be sufficient to point the problem out.

Conclusions

The discussion began with the observation that variation in the length of documents in electronic text corpora can be a problem for a range of interpretative technologies, and undertook to address that problem with reference to exploratory multivariate analysis of frequency data. The discussion was in four main parts. The first part stated the nature of the problem, the second described some normalization methods designed to mitigate or eliminate it, the third identified poor estimation of variable population probability as a factor that compromises the effectiveness of the normalization methods for very short documents, and the fourth proposed elimination of data matrix rows representing document which are too short to be reliably normalized and suggested a way of identifying the relevant documents.

References

[1] J. Bartlett, J. Kotrlik, C. Higgins. Organizational Research: Determining appropriate sample size in survey research. *Information Technology, Learning, and Performance Journal* 19:43-50, 2001.

- [2] C. Buckley. The importance of proper weighting methods. In: M. Bates (ed.). *Human Language Technology*. San Mateo, CA: Morgan Kaufmann, 1993.
- [3] W. Cochran. *Sampling Techniques*, 3rd ed. New York: John Wiley & Sons, 1977.
- [4] *Corpus of Middle English Prose and Verse* (25 February, 2008). <http://quod.lib.umich.edu/c/cme/>.
- [5] R. Czaja, J. Blair. *Designing Surveys. A Guide to Decisions and Procedures*. Thousand Oaks CA: Pine Forge Press, 2005.
- [6] P. Dattalo. *Determining Sample Size*. Oxford: Oxford University Press, 2008.
- [7] B. Everitt, S. Landau, M. Leese.. *Cluster Analysis*, 4th ed. London: Arnold., 2001.
- [8] J. Fraleigh, R. Beauregard. *Linear Algebra*. 2nd ed. Menlo Park, CA: Addison-Wesley, 1995.
- [9] C. Grinstead, J. Snell. *Introduction to Probability*, 2nd ed. American Mathematical Society, 1997.
- [10] J. Hair, W. Black, B. Babin, R. Anderson, R. Tatham. *Multivariate Data Analysis*, 6th ed. New Jersey: Prentice-Hall, 2005.
- [11] I. Jolliffe. *Principal Component Analysis*, 2nd ed. Berlin and Heidelberg: Springer Verlag, 2002.
- [12] P. Levy, S. Lemeshow. *Sampling of Populations. Methods and Applications*, 2nd ed. New York: John Wiley & Sons, 1991.
- [13] J. Milton, J. Arnold. *Introduction to Probability and Statistics*, 4th ed. Boston: McGraw-Hill, 2003.
- [14] *Project Gutenberg* (25 February, 2008). http://www.gutenberg.org/wiki/Main_Page
- [15] A. Singhal, G. Salton, M. Mitra, C. Buckley. Document Length Normalization. *Information Processing and Management* 32: 619-633, 1996.
- [16] A. Singhal, C. Buckley, M. Mitra. Pivoted document length normalization. *Proceedings of the 19th ACM Conference on Research and Development in Information Retrieval (SIGIR-96)*, 21-29, 1996.
- [17] B. Tabachnik, L. Fidell. *Using Multivariate Statistics*, 5th ed.. Boston: Allyn & Bacon, 2006.

