Statistical corpus exploitation

Hermann Moisl, Newcastle University

Abstract

The aim of this chapter is to encourage corpus linguists to use quantitative and more specifically statistical methods in analyzing large digital electronic corpora, focussing in particular on cluster analysis. The first part of the discussion motivates the use of cluster analysis in corpus linguistics, the second gives an outline account of data creation and clustering with reference to the *Newcastle Electronic Corpus of Tyneside English*, and the third is a selective literature review.

Keywords

Quantitative methods, statistics, cluster analysis, data creation, corpus linguistics, Newcastle Electronic Corpus of Tyneside English

1. **Introduction**

This chapter regards corpus linguistics (Kennedy 1998; McEnery & Wilson 2001; Baker 2009) as a methodology for creating collections of natural language speech and text, abstracting data from them, and analysing that data with the aim of generating or testing hypotheses about the structure of language and its use in the world. On this definition, corpus linguistics began in the late eighteenth century with the postulation of an Indo-European protolanguage and its reconstruction based on examination of numerous living languages and of historical texts (Clackson 2007). Since then it has been applied to research across the range of linguistics subdisciplines and, in recent years, has become an academic discipline with its own research community and scientific apparatus of professional organizations, websites, conferences, journals, and textbooks.

Throughout the nineteenth and much of the twentieth centuries corpus linguistics has been mainly or exclusively paper-based. The linguistic material used by researchers was in the form of hand-written or printed documents, and analysis involved reading through the documents, often repeatedly, creating data by noting features of interest on some paper medium such as index cards, inspecting the data directly, and on the basis of that inspection drawing conclusions that were published in printed books or journals. The advent of digital electronic technology in the second half of the twentieth century and its evolution since then have rendered this traditional technology increasingly obsolete. One the one hand, the possibility of representing language electronically rather than as visual marks on paper or some other physical medium, together with the development of digital media and infrastructure and of computational tools for creation, emendation, storage, and transmission of electronic text have led to a rapid increase in the number and size of corpora available to the linguist, and these are now at or even beyond the limit of what an individual researcher can efficiently use in the traditional way. On the other, data abstracted from large corpora can themselves be so extensive and complex as to be impenetrable to understanding by direct inspection. Digital electronic technology has, in general, been a boon to corpus linguistics, but, as with other aspects of life, it's possible to have too much of a good thing.

One response to digital electronic language and data overload is to use only corpora of tractable size or, equivalently, subsets of large corpora, but simply ignoring available information is not scientifically respectable. The alternative is to look to related research disciplines for help. The overload in corpus linguistics is symptomatic of a more general trend. Daily use of digital electronic information technology by many millions of people worldwide both in their professional and personal lives has generated and continues to generate truly vast amounts of electronic speech and text, and abstraction of information from all but a tiny part of it by direct inspection is an intractable task not only for individuals but also in government an commerce -- what, for example, are the prospects for finding a specific item of information by reading sequentially through the huge number of documents currently available on the Web? In response, research disciplines devoted to information abstraction from very large collections of electronic text have come into being, among them Computational Linguistics (Mitkov 2005), Natural Language Processing (Manning & Schütze 1999; Dale et al. 2000; Jurafsky & Martin 2008; Cole et al.2010; Indurkhya & Damerau 2010), Information Retrieval (Manning et al. 2008), and Data Mining (Hand et al. 2001). These disciplines use existing statistical methods supplemented by a range of new interpretative ones to develop tools that render the deluge of digital electronic text tractable. Many of these methods and tools are readily adaptable for corpus linguistics use, and, as the references in section 3 below demonstrate, interest in them has grown substantially in recent years. The general aim of this chapter is to encourage that growth, and the particular aim is to encourage it with respect to corpus-based phonetic and phonological research.

The chapter is in three main parts. The first part motivates the selection of one particular class of statistical method, cluster analysis, as the focus of the discussion, the second describes fundamental concepts in cluster analysis and exemplifies their application to hypothesis generation in corpus-based phonetic and phonological research, and the third reviews the literature on the use of statistical methods in general and of cluster analysis more specifically in corpus linguistics.

2. **Cluster analysis: motivation**

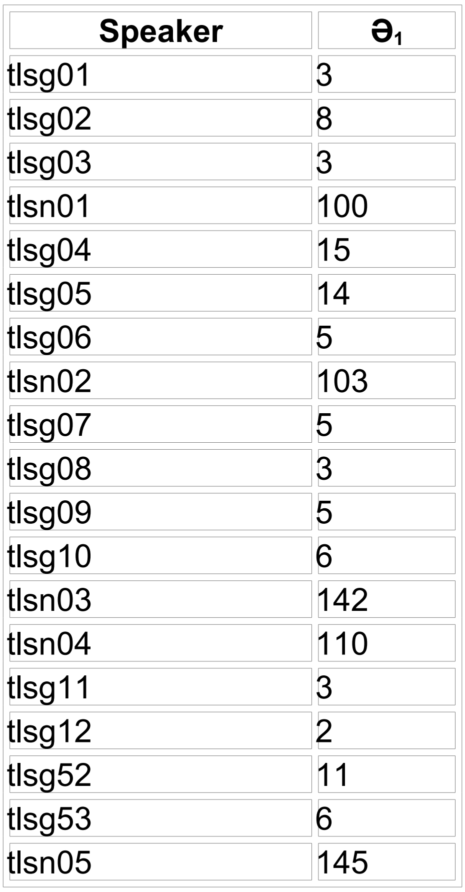
'Statistics' encompasses an extensive range of mathematical concepts and techniques with a common focus: understanding of the nature of probability and of its role in the behaviour of natural systems. Linguistically-oriented statistical analysis of a natural language corpus thus implies that the aim of the analysis is in some sense to interpret the probabilities of occurrence of one or more features of interest -- phonetic, phonological, morphological, lexical, syntactic, or semantic-- in relation to some research question.

The statistics literature makes a fundamental distinction between exploratory and confirmatory analysis. Confirmatory analysis is used when the researcher has formulated a hypothesis in answer to his or her research question about a domain of interest, and wants to test the validity of that hypothesis by abstracting data from a sample drawn from the domain and applying confirmatory statistical methods to those data. Exploratory analysis is, on the other hand, used when the researcher has not yet formulated a hypothesis and wishes to generate one by abstracting data from a sample of the domain and then looking for structure in the data on the basis of which a reasonable hypothesis can be formulated. The present discussion purports to describe statistical corpus exploitation, and as such it should cover both these types of analysis. The range of material which this implies is, however, very extensive, and attempting to deal even with only core topics in a relatively short chapter would necessarily result in a sequence of tersely described abstract concepts with little or no discussion of their application to corpus analysis. Since the general aim is to encourage rather than to discourage, some selectivity of coverage is required.

The selection of material for discussion was motivated by the following question: given the proliferation of digital electronic corpora referred to in the Introduction, which statistical concepts and techniques would be most useful to corpus linguists for dealing with the attendant problem of analytical intractability? The answer was exploratory rather than confirmatory analysis. The latter is appropriate where the characteristics of the domain of interest are sufficiently well understood to permit formulation of sensible hypotheses; in corpus linguistic terms such a domain might be a collection of texts in the English language, which has been intensively studied, or one that is small enough to be tractable by direct inspection. Where the corpora are very large, however, or in languages / dialectal varieties that are relatively poorly understood, or both, exploratory analysis is more useful because it provides a basis for the formulation of reasonable hypotheses; such hypotheses can subsequently be tested using confirmatory methods.

The range of exploratory methods is itself extensive (Myatt 2006; Myatt & Johnson 2009), and further restriction is required. To achieve this, a type of problem that can be expected to occur frequently in exploratory corpus analysis was selected and the relevant class of analytical methods made the focus of the discussion. Corpus exploration implies some degree of uncertainty about what one is looking for. If, for example, the aim is to differentiate the documents in a collection on the basis of their lexical semantic content, which words are the best differentiating criteria? Or, if the aim is to group a collection of speaker interviews on the basis of their phonetic characteristics, which phonetic features are most important? In both cases one would want to take as many lexical / phonetic features as possible into account initially, and then attempt to identify the important ones among them in the course of exploration. Cluster analysis is a type of exploratory method that has long been used across a wide range of science and engineering disciplines to address this type of problem, and is the focus of subsequent discussion. The remainder of this section gives an impression of what cluster analysis involves and how it can be applied to corpus analysis; a more detailed account is given in section 2.

Observation of nature plays a fundamental role in science. But nature is dauntingly complex, and there is no practical or indeed theoretical hope of describing any aspect of it objectively and exhaustively. The researcher is therefore selective in what he or she observes: a research question about the domain of interest is posed, a set of variables descriptive of the domain in relation to the research question is defined, and a series of observations is conducted in which, at each observation, the quantitative or qualitative values of each variable are recorded. A body of data is therefore built up on the basis of which a hypothesis can be generated. Say, for example, that the domain of interest is the phonetic usage of the speakers in some corpus, and that the research question is whether there is any systematic variation in phonetic usage among the speakers. Figure 1 shows data abstracted from the *Newcastle Electronic Corpus of Tyneside English* (NECTE) (Allen et al. 2007), a corpus of dialect speech from north-east England which is described in chapter IV.5 of this volume.



Speaker	**ə$_1$**
tlsg01	3
tlsg02	8
tlsg03	3
tlsn01	100
tlsg04	15
tlsg05	14
tlsg06	5
tlsn02	103
tlsg07	5
tlsg08	3
tlsg09	5
tlsg10	6
tlsn03	142
tlsn04	110
tlsg11	3
tlsg12	2
tlsg52	11
tlsg53	6
tlsn05	145

tlsn06	109
tlsg54	3
tlsg55	7
tlsg56	12
tlsn07	104

Figure 1: Frequency data for Ə$_1$ in the NECTE corpus

The speakers are described by a single variable, the phonetic segment Ə$_1$, and the values in the variable column of figure 1 are the frequencies with which each of the 24 speakers use that segment. It is easy to see by direct inspection that the speakers fall into two groups: those that use Ə$_1$ relatively frequently, and those that use it relatively infrequently. The hypothesis is, therefore, that there is systematic variation in phonetic usage among NECTE speakers. If two phonetic variables are used to describe the speakers, as in figure 2, direct inspection again shows two groups, those that use both Ə$_1$ and Ə$_2$ relatively frequently and those that do not, and the hypothesis remains the same.

Speaker	**Ə$_1$**	**Ə$_2$**
tlsg01	3	1
tlsg02	8	0
tlsg03	3	1
tlsn01	100	116
tlsg04	15	0
tlsg05	14	6
tlsg06	5	0
tlsn02	103	93
tlsg07	5	0
tlsg08	3	0
tlsg09	5	0
tlsg10	6	0
tlsn03	142	107
tlsn04	110	120

tlsg11	3	0
tlsg12	2	0
tlsg52	11	1
tlsg53	6	0
tlsn05	145	102
tlsn06	109	107
tlsg54	3	0
tlsg55	7	0
tlsg56	12	0
tlsn07	104	93

Figure 2: Frequency data for $Ə_1$ and $Ə_2$ in the NECTE corpus

There is no theoretical limit on the number of variables that can be defined to describe the objects in a domain. As the number of variables and observations grows, so does the difficulty of generating hypotheses from direct inspection of the data. In the NECTE case, the selection of $Ə_1$ and $Ə_2$ in figures 1 and 2 was arbitrary, and the speakers could be described using more phonetic segment variables. Figure 3 shows twelve.

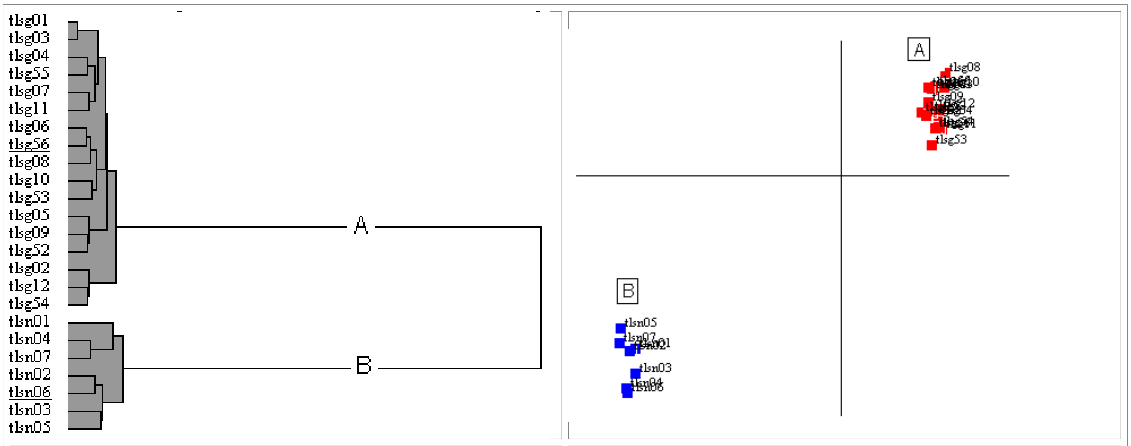
Speaker	**$ə_1$**	**$ə_2$**	**o:**	**$ə_3$**	**ī**	**eī**	**n**	**$a:_1$**	**$a:_2$**	**aī**	**r**	**w**
tlsg01	3	1	55	101	33	26	193	64	1	8	54	96
tlsg02	8	0	11	82	31	44	205	54	64	8	83	88
tlsg03	3	1	55	101	33	26	193	64	15	8	54	96
tlsn01	100	116	5	17	75	0	179	64	0	19	46	62
tlsg04	15	0	12	75	21	23	186	57	6	12	32	97
tlsg05	14	6	45	70	49	0	188	40	0	45	72	79
tlsg06	5	0	40	70	32	22	183	46	0	2	37	117
tlsn02	103	93	7	5	87	27	241	52	0	1	19	72
tlsg07	5	0	11	58	44	31	195	87	12	4	28	93
tlsg08	3	0	44	63	31	44	140	47	0	5	43	106
tlsg09	5	0	30	103	68	10	177	35	0	33	52	96
tlsg10	6	0	89	61	20	33	177	37	0	4	63	97
tlsn03	142	107	2	15	94	0	234	15	0	25	28	118
tlsn04	110	120	0	21	100	0	237	4	0	61	21	62
tlsg11	3	0	61	55	27	19	205	88	0	4	47	94
tlsg12	2	0	9	42	43	41	213	39	31	5	68	124

tlsg52	11	1	29	75	34	22	206	46	0	29	34	93
tlsg53	6	0	49	66	41	32	177	52	9	1	68	74
tlsn05	145	102	4	6	100	0	208	51	0	22	61	104
tlsn06	109	107	0	7	111	0	220	38	0	26	19	70
tlsg54	3	0	8	81	22	27	239	30	32	8	80	116
tlsg55	7	0	12	57	37	20	187	77	41	4	58	101
tlsg56	12	0	21	59	31	40	164	52	17	6	45	103
tlsn07	104	93	0	11	108	0	194	5	0	66	33	69

Figure 3: Frequency data for a range of phonetic segments in the NECTE corpus

What hypothesis would one formulate from inspection of the data in figure 3, taking into account all the variables? There are, moreover, 64 speakers in the NECTE corpus and the transcription scheme contains 156 phonetic segments, so it is possible to describe the phonetic usage of each or 64 speakers in terms of 156 variables. What hypothesis would one formulate from direct inspection the full 64 x 156 data? These questions are clearly rhetorical, and there is a straightforward moral: human cognitive makeup is unsuited to seeing regularities in anything but the smallest collections of numerical data. To see the regularities we need help, and that is what cluster analysis provides.

Cluster analysis is a family of mathematical methods for identification and graphical display of structure in data when the data is too large either in terms of the number of variables or of the number of objects described, or both, for it to be readily interpretable by direct inspection. All the members of the family work by partitioning a set of objects in the domain of interest into disjoint subsets in accordance with how relatively similar those objects are in terms of the variables that describe them. The objects of interest in the NECTE data are speakers, and each speaker's phonetic usage is described by a set of phonetic variables. Any two speakers' phonetic usage will be more or less similar depending on how similar their respective variable values are: if the values are identical then so are the speakers in terms of their phonetics, and the greater the divergence in values the greater the differences in usage. Cluster analysis of the NECTE data in figure 3 groups the 24 speakers in terms of how similar their frequency of usage of 12 of the full 156 phonetic segments is. There are various kinds of cluster analysis; figure 4 shows the results from application of two of them.



a: Hierarchical cluster analysis of the NECTE data

b: Multidimensional scaling analysis of the NECTE data

Figure 4: Two types of cluster analysis of the data in figure 3

Figure 4a shows the cluster structure of the NECTE data in figure 3 as a hierarchical tree. To interpret the tree one has to understand how it is constructed, so a short intuitive account is given here; technical details are given later in the discussion. The labels at the leaves of the tree are the speaker-identifiers. These labels are partitioned into clusters in a sequence of steps. Initially, each speaker is interpreted as a cluster on his or her own. At the first step the data is searched to identify the two most similar clusters. When found, they are joined into a superordinate tree in which their degree of similarity is graphically represented as the length of the horizontal lines joining the subclusters: the more similar the subclusters, the shorter the lines. In the actual clustering procedure assessment of similarity is done numerically, but for present expository purposes visual inspection of figure 4a is sufficient, and, to judge by the shortness of the horizontal lines, the singleton clusters tlsg01 and tlsg03 at the top of the tree are the most similar. These are joined into a composite cluster (tlsg01 tlsg03). At the second step the data is searched again to determine the next-most-similar pair of clusters. Visual inspection indicates that these are tlsg06 and tlsg56 about 1/3 of the way down the tree, and these are joined into a composite cluster (tlsg06 tlsg56). At step 3, the two most similar clusters are the composite cluster (tlsg06 tlsg56) constructed at step 2 and tlsg08. These are joined into a superordinate cluster ((tlsg06 tlsg56) tlsg08). The sequence of steps continues in this way, combining the most similar pair of clusters at each step, and stops when there is only one cluster remaining which contains all the subclusters. The resulting tree gives an exhaustive graphical representation of the similarity relations in the NECTE speaker data. It shows that there are two main groups of speakers, labelled A and B, which differ greatly from one another in terms of phonetic usage, and, though there are differences in usage among the speakers in those two main groups, the differences are minor relative to those between A and B.

Figure 4b shows the cluster structure of the data in figure 3 as a scatter plot in which relative spatial distance between speaker labels represents the relative similarity of phonetic usage among the speakers: the closer the labels the closer the speakers. Labels corresponding to the main clusters in figure 4a have been added for ease of cross-reference, and show that this analysis gives the same result as the hierarchical one.

Once the structure of the data has been identified by cluster analysis, it can be used for hypothesis generation (Romesburg 1984, chs. 4 & 22). The obvious hypothesis in the present case is that the NECTE speakers fall into two distinct groups in terms of their phonetic usage. This could be tested by doing an analysis of the full NECTE corpus using all 64 speakers and all 156 variables, and by conducting further interviews and abstracting data from them for subsequent analysis.

Cluster analysis can be applied in any research where the data consists of objects described by variables; since most research uses data of this kind, cluster analysis is very widely applicable. It can *usefully* be applied where the number of objects and descriptive variables is so large that the data cannot easily be interpreted by direct inspection, and the range of applications where this is the case spans most areas of science, engineering, and commerce (Everitt et al. 2011, ch. 1; Romesburg 1984, chs. 4-6; detailed discussion of cluster applications in Jain et al. 1999, 296 ff). In view of the comments made in the Introduction about text overload, cluster analysis is exactly what's required for hypothesis generation in corpus linguistics. The foregoing discussion of NECTE is an example in the intersection of phonetics, dialectology, and sociolinguistics: the set of phonetic transcriptions is extensive and the frequency data abstracted from them is far too large to be in any sense comprehensible, but the structure that cluster analysis identified in the data made hypothesis formulation straightforward.

3 **Cluster analysis concepts and hypothesis generation**

3.1 ***Data***

Data are abstractions of what we observe using our senses, often with the aid of instruments (Chalmers 1999), and are ontologically different from the world. The world is as it is; data are an interpretation of it for the purpose of scientific study. The weather is not the meteorologist's data –measurements of such things as air temperature are. A text corpus is not the linguist's data – measurements of such things as word frequency are. Data are constructed from observation of things in the world, and the process of construction raises a range of issues that determine the amenability of the data to analysis and the interpretability of the results. The importance of understanding such data issues in cluster analysis can hardly be overstated. On the one hand, nothing can be discovered that is beyond the limits of the data itself. On the other, failure to understand relevant characteristics of data can lead to results and interpretations that are distorted or even worthless. For these reasons, an overview of data issues is given before moving on to discussion of cluster analysis concepts; examples are taken from the NECTE corpus cited above.

3.1.1 *Formulation of a research question*

In general, any aspect of the world can be described in an arbitrary number of ways and to arbitrary degrees of precision. The implications of this go straight to the heart of the debate on the nature of science and scientific theories, but to avoid being drawn into that debate, this discussion adopts the position that is pretty much standard in scientific practice: the view, based on Karl Popper's philosophy of science (Popper 1959; Popper 1963; Chalmers 1999), that there is no theory-free observation of the world. In essence, this means that there is no such thing as objective observation in science. Entities in a domain of inquiry only become relevant to observation in terms of a research question framed using the ontology and axioms of a theory about the domain. For example, in linguistic analysis, variables are selected in terms of the discipline of linguistics broadly defined, which includes the division into subdisciplines such as sociolinguistics and dialectology, the subcategorization within subdisciplines such as phonetics through syntax to semantics and pragmatics in formal grammar, and theoretical entities within each subcategory such as phonemes in phonology and constituency structures in syntax. Claims, occasionally seen, that the variables used to describe a corpus are 'theoretically neutral' are naive: even word categories like 'noun' and 'verb' are interpretative constructs that imply a certain view of how language works, and they only appear to be theory-neutral because of familiarity with long-established tradition. Data can, therefore, only be created in relation to a research question that is defined using the ontology of the domain of interest, and that thereby provides an interpretative orientation. Without such an orientation, how does one know what to observe, what is important, and what is not? The research question asked with respect to the NECTE corpus, and which serves as the basis for the examples in what follows, is:

> *Is there systematic phonetic variation in the Tyneside speech community, and , if so, what are the main phonetic determinants of that variation?*

3.1.2 *Variable selection*

Given that data are an interpretation of some domain of interest, what does such an interpretation look like? It is a description of entities in the domain in terms of variables. A variable is a symbol, and as such is a physical entity with a conventional semantics, where a conventional semantics is understood as one in which the designation of a physical thing as a symbol together with the connection between the symbol and what it represents are determined by agreement within a community. The symbol ‘A’, for example, represents the phoneme /a/ by common assent, not because there is any necessary connection between it and what it represents. Since each variable has a conventional semantics, the set of variables chosen to describe entities constitutes the template in terms of which the domain is interpreted. Selection of appropriate variables is, therefore, crucial to the success of any data analysis.

Which variables are appropriate in any given case? That depends on the nature of the research question. The fundamental principle in variable selection is that the variables must describe all and only those aspects of the domain that are relevant to the research question. In general, this is an unattainable ideal. Any domain can be described by an essentially arbitrary number of finite sets of variables; selection of one particular set can only be done on the basis of personal knowledge of the domain and of the body of scientific theory associated with it, tempered by personal discretion. In other words, there is no algorithm for choosing an optimally relevant set of variables for a research question.

The NECTE speakers are described by a set of 156 variables each of which represents a phonetic segment. These are described in (Allen et al. 2007) and, briefly, in chapter IV.5 of this volume.

3.1.3 *Variable value assignment*

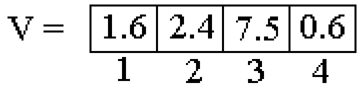
The semantics of each variable determines a particular interpretation of the domain of interest, and the domain is 'measured' in terms of the semantics. That measurement constitutes the values of the variables: height in metres = 1.71, weight in kilograms = 70, and so on. Measurement is fundamental in the creation of data because it makes the link between data and the world, and thus allows the results of data analysis to be applied to the understanding of the world.

Measurement is only possible in terms of some scale. There are various types of measurement scale, and these are discussed at length in any statistics textbook, but for present purposes the main dichotomy is between numeric and non-numeric. Cluster analysis methods assume numeric measurement as the default case, and for that reason the same is assumed in what follows. For NECTE we are interested in the number of times each speaker uses each of the phonetic segment variables. The speakers are therefore 'measured' in terms of the frequency with which they use these segments

3.1.4 *Data representation*

If they are to be analyzed using mathematically-based computational methods, the descriptions of the entities in the domain of interest in terms of the selected variables must be mathematically represented. A widely used way of doing this, and the one adopted here, is to use structures from a branch of mathematics known as linear algebra. There are numerous textbooks and websites devoted to linear algebra; a small selection of introductory textbooks is (Fraleigh & Beauregard 1994; Poole 2005; Strang 2009).

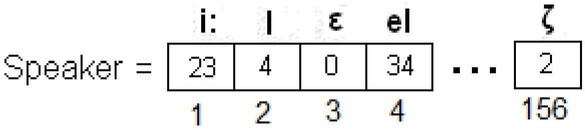
Vectors are fundamental in data representation. A vector is a sequence of numbered slots containing numerical values. Figure 5 shows a four-element vector each element of which contains a real-valued number: 1.6 is the value of the first element v_1, 2.4 the value of the second element v_2, and so on.



V =	1.6	2.4	7.5	0.6
	1	2	3	4

Figure 5: A vector

A single NECTE speaker's frequency of usage of the 156 phonetic segments in the transcription scheme can be represented by a 156-element vector in which each element is associated with a different segment, as in figure 6.

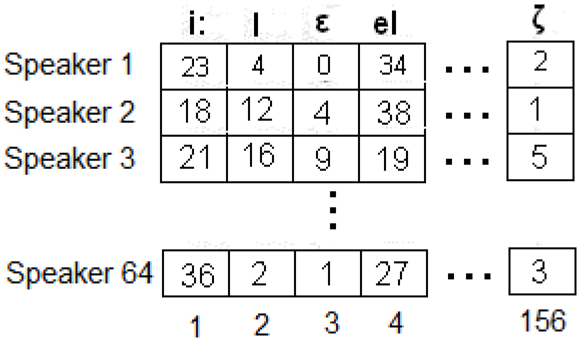


	i:	I	ε	eI		ζ
Speaker =	23	4	0	34	. . .	2
	1	2	3	4		156

Figure 6: A vector representing a NECTE speaker

This speaker uses the segment at Speaker_1 twenty three times, the segment at Speaker_2 four times, and so on.

The 64 NECTE speaker vectors can be assembled into a matrix M, shown in figure 7, in which the 64 rows represent the speakers, the 156 columns represent the phonetic segments, and the value at M_{ij} is the number of times speaker *i* uses segment *j* (for i = 1..64 and j = 1..156):



	i:	I	ε	eI		ζ
Speaker 1	23	4	0	34	. . .	2
Speaker 2	18	12	4	38	. . .	1
Speaker 3	21	16	9	19	. . .	5
			⋮			
Speaker 64	36	2	1	27	. . .	3
	1	2	3	4		156

Figure 7: The NECTE data matrix

This matrix M is the basis of subsequent examples.

3.1.5 *Data issues*

Once the data is in matrix form it can in principle be cluster analyzed. It may, however, have characteristics that can distort or even invalidate the results, and any such characteristics have to be mitigated or eliminated prior to analysis. These include variation in document or speaker interview length

(Moisl 2009), differences in variable measurement scale (Moisl 2010), data sparsity (Moisl 2008), and nonlinearity (Moisl 2007).

3.2 ***Cluster analysis***

Once the data matrix has been created and any data issues resolved, a variety of computational methods can be used to group its row vectors, and thereby the objects in the domain that the row vectors represent. In the present case, those objects are the NECTE speakers.

3.2.1 *Clusters in vector space*

Though it is just a sequence of numbers, a vector can be geometrically interpreted (Fraleigh & Beauregard 1994; Poole 2005; Strang 2009). To see how, take a vector consisting of two elements, say *v* = (30,70). Under a geometrical interpretation, the two elements of *v* define a two-dimensional space, the numbers at v_1 = 30 and v_2 = 70 are coordinates in that space, and the vector *v* itself is a point at the coordinates (30,70), as shown in figure 8.

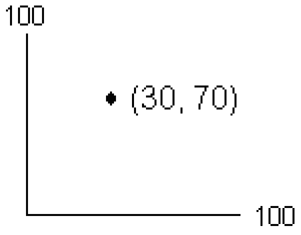


Figure 8: Geometrical interpretation of a 2-dimensional vector

A vector consisting of three elements, say *v* = (40, 20, 60) defines a three-dimensional space in which the coordinates of the point v are 40 along the horizontal axis, 20 along the vertical axis, and 60 along the third axis shown in perspective, as in figure 9.

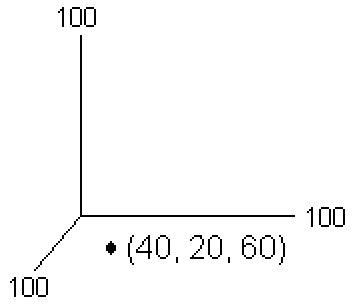


Figure 9: Geometrical interpretation of a 3-dimensional vector

A vector *v* = (22, 38, 52, 12) defines a four-dimensional space with a point at the stated coordinates, and so on to any dimensionality *n*. Vector spaces of dimensionality greater than 3 are impossible to visualize directly and are therefore counterintuitive, but mathematically there is no problem with them; two and three dimensional spaces are useful as a metaphor for conceptualizing higher-dimensional ones.

When numerous vectors exist in a space, it may or may not be possible to see interesting structure in the way they are arranged in it. Figure 10 shows vectors in two and three dimensional spaces. In (a) they were randomly generated and there is no structure to be observed, in (b) there are two clearly defined concentrations in two dimensional space, and in (c) there are two clearly defined concentrations in three-dimensional space.

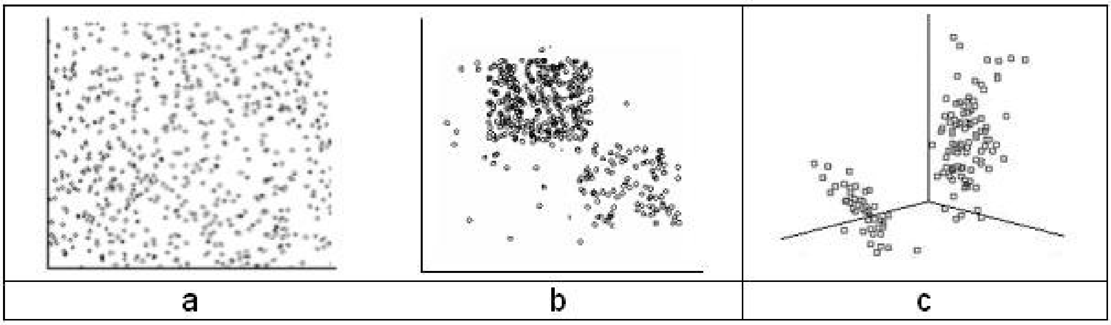


Figure 10: Distributions of multiple vectors in 2 and 3-dimensional spaces

The existence of concentrations like those in (b) and (c) indicate relationships among the entities that the vectors represent. In (b), for example, if the horizontal axis measures weight and the vertical one height for a sample human population, then members of the sample fall into two groups: tall, light people on the one hand, and short heavy ones on the other.

This idea of identifying clusters of vectors in vector space and interpreting them in terms of what the vectors represent is the basis of cluster analysis. In what follows, we shall be attempting to group the NECTE speakers on the basis of their phonetic usage by looking for clusters in the arrangement of the row vectors of M in 156-dimensional space.

3.2.2 *Clustering methods*

Where the data vectors are two or three-dimensional they can simply be plotted and any clusters will be visually identifiable, as we have just seen. But what about when the vector dimensionality is greater than 3 -say 4, or 10, or 100? In such a case direct plotting is not an option; how exactly would one draw a 6-dimensional space, for example? Many data matrix row vectors have dimensionalities greater than 3 --the NECTE matrix M has dimensionality

156-- and, to identify clusters in such high-dimensional spaces some procedure more general than direct plotting is required. A variety of such procedures is available, and they are generically known as cluster analysis methods. This section looks at these methods.

Where there are two or more vectors in a space, it is possible to measure the distance between any two of them and to rank them in terms of their proximity to one another. Figure 11 shows a simple case of a 2-dimensional space in which the distance from vector A to vector B is greater than the distance from A to C.

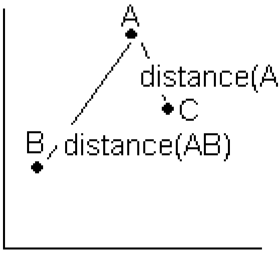


Figure 11: Vector distances

There are various ways of measuring such distances, but the most often used is the familiar Euclidean one, as in figure 12:

$$dist(AB) = \sqrt{(5-1)^2 + (4-2)^2}$$

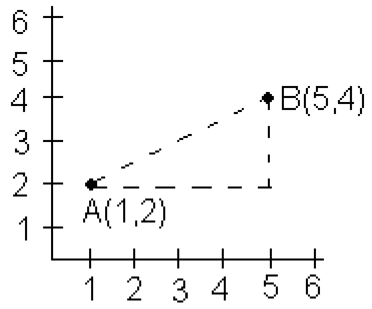
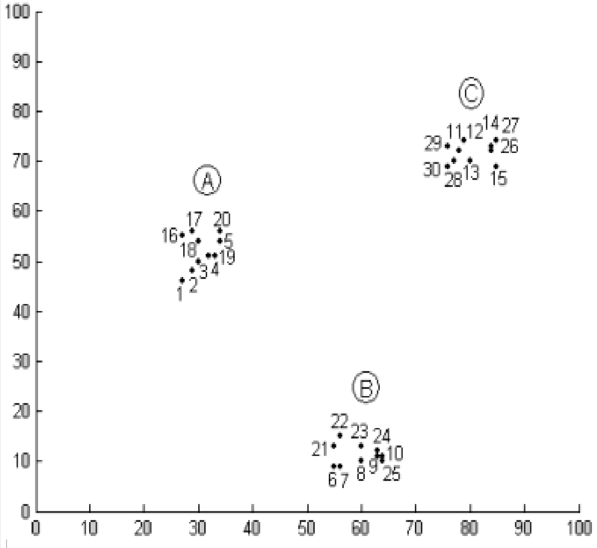


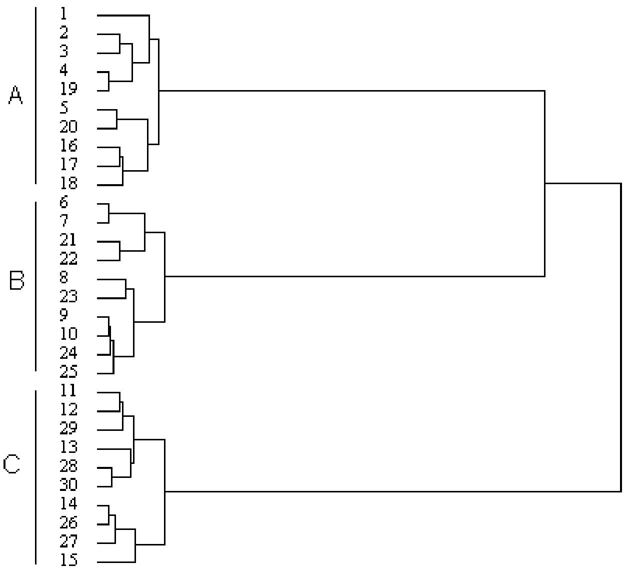
Figure 12: Euclidean distance measurement

Cluster analysis methods use relative distance among vectors in a space to group the vectors. Specifically, for a given set of vectors in a space, they first calculate the distances between all pairs of vectors, and then group into clusters all the vectors that are relatively close to one another in the space and relatively far from those in other clusters. 'Relatively close' and 'relatively far' are, of course, vague expressions, but they are precisely defined by the various clustering methods, and for present purposes we can avoid the technicalities and rely on intuitions about relative distance.

For concreteness, we will concentrate on one particular class of methods: the hierarchical cluster analysis already introduced in section 1 above, which represents the relativities of distance among vectors as a tree. Figure 13 exemplifies this.

	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

a	b

Figure 13: Data matrix and corresponding row-clusters

Column (a) shows a 30 x 2 data matrix that is to be cluster analyzed. Because the data space is 2-dimensional the vectors can be directly plotted to show the cluster structure, as in the upper part of column (b). The corresponding hierarchical cluster tree is shown in the lower part of column (b). There are three clusters labelled A, B, and C in each of which the distances among vectors are quite small. These three clusters are relatively far from one another, though A and B are closer to one another than either of them is to C. Comparison with the vector plot shows that the hierarchical analysis accurately represents the distance relations among the 30 vectors in 2-dimensional space.

Given that the tree tells us nothing more than what the plot tells us, what is gained? In the present case, nothing. The real power of hierarchical analysis lies in its independence of vector space dimensionality. We have seen that direct plotting is limited to three or fewer dimensions, but there is no dimensionality limit on hierarchical analysis -it can determine relative distances in vector spaces of any dimensionality and represent those distance relativities as a tree like the one above. To exemplify this, the 156-dimensional NECTE data matrix M was hierarchically cluster analyzed (Moisl et al. 2006), and the result is shown in figure 14.

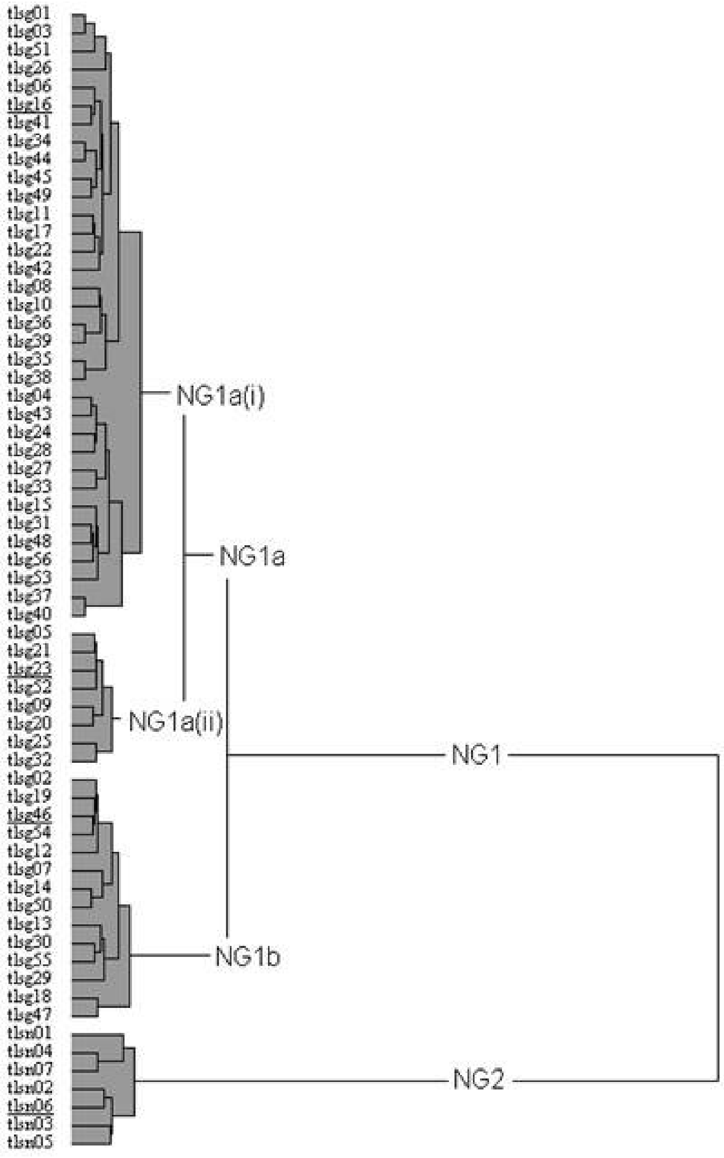


Figure 14: Hierarchical analysis of the NECTE data matrix in figure 5

Plotting M in 156-dimensional space would have been impossible, and, without cluster analysis, one would have been left pondering a very large and incomprehensible matrix of numbers. With the aid of cluster analysis, however, structure in the data is clearly visible: there are two main clusters, NG1 and NG2; NG1 consists of large subclusters NG1a and NG1b; NG1a itself has two main subclusters NG1a(i) and NG1a(ii).

3.2.3 *Hypothesis generation*

Given that there is structure in the relative distances of the row vectors of M from one another in the data space, what does that structure mean in terms of the research question?

> '*Is there systematic phonetic variation in the Tyneside speech community, and, if so, what are the main phonetic determinants of that variation?*'

Because the row vectors of M are phonetic profiles of the NECTE speakers, the cluster structure means that the speakers fall into clearly defined groups with specific interrelationships rather than, say, being randomly distributed around the phonetic space. A reasonable hypothesis to answer the first part of the research question, therefore, is that there is systematic variation in the Tyneside speech community. This hypothesis can be refined by examining the social data relating to the NECTE speakers, which shows, for example, that all those in the NG1 cluster come from the Gateshead area on the south side of the river Tyne and all those in NG2 come from Newcastle on the north side, and that the subclusters in NG1 group the Gateshead speakers by gender and occupation (Moisl et al. 2006).

The cluster tree can also be used to generate a hypothesis in answer to the second part of the research question. So far we know *that* the NECTE speakers fall into clearly-demarcated groups on the basis of variation in their phonetic usage. We do not, however, know *why,* that is, which segments out of the 156 in the TLS transcription scheme are the main determinants of this regularity. To identify these segments (Moisl & Maguire 2008), we begin by looking at the two main clusters NG1 and NG2 to see which segments are most important in distinguishing them.

The first step is to create for the NG1 cluster a vector that captures the general phonetic characteristics of the speakers it contains, and to do the same for the NG2. Such vectors can be created by averaging all the row vectors in a cluster using the formula

$$v_j = \frac{\sum_{i=1..m} M_{ij}}{m}$$

where v_j is the *j*th element of the average or 'centroid' vector *v* (for *j* = 1..the number of columns in M), M is the data matrix, Σ designates summation, and *m* is the number of row vectors in the cluster in question (56 for NG1, 7 for NG2). This yields two centroid vectors.

Next, compare the two centroid vectors by co-plotting them to show graphically how, on average, the two speaker groups differ on each of the 156 phonetic segments; a plot of all 156 segments is too dense to be readily deciphered, so the six on which the NG1 and NG2 centroids differ most are shown in Figure 15.

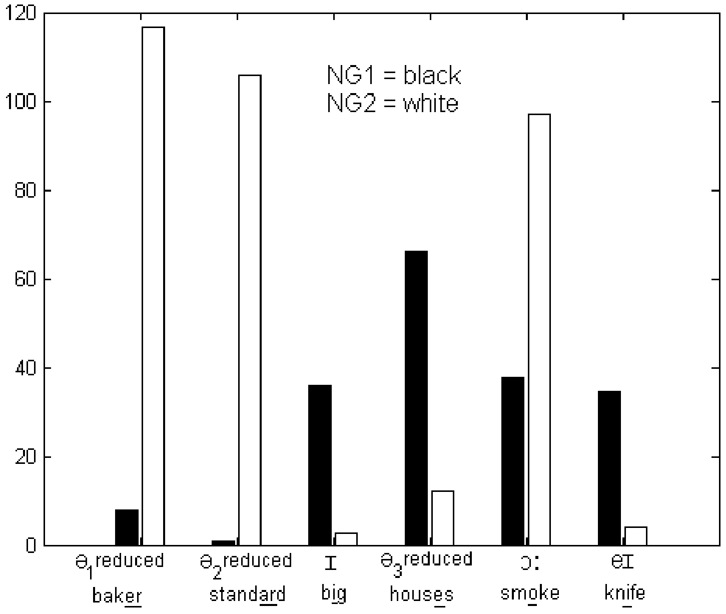


Figure 15: Co-plot of centroid vectors for NG1 and NG2

The six phonetic segments most important in distinguishing cluster NG1 from NG2 are three varieties of (ə), (ɔː), (ɪ), and (eɪ): the Newcastle speakers characteristically use $Ə_1$ and $Ə_2$ whereas the Gateshead speakers use them hardly at all, the Gateshead speakers use $Ə_3$ much more than the Newcastle speakers, and so on. A hypothesis that answers the second part of the research question is therefore that the main determinants of phonetic variation in the Tyneside speech community are three kinds of (ə), (ɔː), (ɪ), and (eɪ).

The subclusters of NG1 can be examined in the same way and the hypothesis thereby further refined.

4. **Literature Review**

The topic of this chapter cuts across several academic disciplines, and the potentially relevant literature is correspondingly large. This review is therefore highly selective. It also includes a few websites; as ever with the Web, *caveat emptor*, but the ones cited seem to me to be reliable and useful.

4.1. Statistics and linear algebra

Using cluster analysis competently requires some knowledge both of statistics and of linear algebra. The following references to introductory and intermediate-level accounts provide this.

4.1.1 Statistics

In any research library there is typically a plethora of introductory and intermediate-level textbooks on probability and statistics. It's difficult to recommend specific ones on a principled basis because most of them, and especially the more recent ones, offer comprehensive and accessible coverage of the fundamental statistical concepts and techniques relevant to corpus analysis. For the linguist at any but advanced level in statistical corpus analysis, choice is usually determined by a combination of what is readily available and presentational style. Some personal introductory favourites are (Devore & Peck 2005; Freedman et al. 2007; Gravetter & Wallnau 2008), and, among more advanced ones, (Casella & Berger 2001; Freedman 2009; Rice 2006).

Statistics websites

- Hyperstat Online Statistics Textbook: http://davidmlane.com/hyperstat/
- NIST-Sematech e-Handbook of Statistical Methods: http://www.itl.nist.gov/div898/handbook/index2.htm
- Engineering Statistics Handbook: http://www.itl.nist.gov/div898/handbook/index.htm
- Statistics on the Web: http://my.execpc.com/~helberg/statistics.html
- Statsoft Electronic Statistics Textbook: http://www.statsoft.com/textbook/
- SticiGui e-textbook: http://www.stat.berkeley.edu/~stark/SticiGui/index.htm
- John C. Pezzullo's Statistical Books, Manuals and Journals links: http://statpages.org/javasta3.html
- Research Methods Knowledge Base: http://www.socialresearchmethods.net/kb/index.php

Statistics software

Contemporary research environments standardly provide one or more statistics packages as part of their IT portfolio, and these packages together with local expertise in their use are the first port of call for the corpus analyst. Beyond this, a Web search using the keywords 'statistics software' generates a deluge of links from which one can choose. Some useful directories are:

- Wikipedia list of statistical software:

  http://en.wikipedia.org/wiki/List_of_statistical_packages

- Wikipedia comparison of statistical packages:

  http://en.wikipedia.org/wiki/Comparison_of_statistical_packages

- Open directory project, statistics software:

  http://www.dmoz.org/Science/Math/Statistics/Software/

- Stata, statistical software providers:

  http://www.stata.com/links/stat_software.html

- Free statistics:

  http://www.freestatistics.info/

- Statcon, annotated list of free statistical software:

  http://statistiksoftware.com/free_software.html

- Understanding the World Today, free software: statistics:

  http://gsociology.icaap.org/methods/soft.html

- The Impoverished Social Scientist's Guide to Free Statistical Software and Resources:

  http://maltman.hmdc.harvard.edu/socsci.shtml

- StatSci, free statistical packages:

  http://www.statsci.org/free.html

- Free statistical software directory:

  http://www.freestatistics.info/stat.php

- John C. Pezullo's free statistical software links:

  http://statpages.org/javasta2.html

- Statlib: http://lib.stat.cmu.edu/

4.1.2 Linear algebra

Much of the literature on linear algebra can appear abstruse to the non-mathematician. Two recent and accessible introductory textbooks are (Poole 2005) and (Strang 2009); older, but still a personal favourite, is (Fraleigh & Beauregard 1994).

Linear algebra websites

- PlanetMath: Linear algebra: http://planetmath.org/encyclopedia/LinearAlgebra.html
- Math Forum: Linear algebra: http://mathforum.org/linear/linear.html

4.2 Cluster analysis

As with general statistics, the literature on cluster analysis is extensive. It is, however, much more difficult to find introductory-level textbooks for cluster analysis, since most assume a reasonable mathematical competence. A good place to start is with (Romesburg 1984), a book that is now quite old but still a standard introductory text. More advanced accounts, in chronological order, are (Jain & Dubes 1988; Arabie et al. 1996; Gordon 1999; Jain et al. 1999; Kaufman & Rousseeuw 2005; Gan et al. 2007; Xu & Wunsch 2008; Everitt et al. 2011). Cluster analysis is also covered in textbooks for related disciplines, chief among them multivariate statistics (Kachigan 1991; Grimm & Yarnold

2000; Hair et al. 2007; Härdle & Simar 2007), data mining (Mirkin 2005; Nisbet et al. 2009), and information retrieval (Manning et al. 2008).

Cluster analysis websites

- Berkhin, P. (2002) Survey of clustering data mining techniques: http://citeseer.ist.psu.edu/cache/papers/cs/26278/http:zSzzSzwww.accrue.comzSzproductszSzrp_cluster_review.pdf/berkhin02survey.pdf
- Journal of Classification: http://www.springer.com/statistics/statistical+theory+and+methods/journal/357

Cluster analysis software

Many general statistics packages provide at least some cluster analytical functionality. For clustering-specific software a Web search using the keywords 'clustering software' or 'cluster analysis sofware' generates numerous links. See also the following directories:

- Classification Society of North America, cluster analysis software: http://www.pitt.edu/~csna/software.html
- Statlib: http://lib.stat.cmu.edu/
- Open Directory Project, cluster analysis: http://search.dmoz.org/cgi-bin/search?search=cluster+analysis&all=yes&cs=UTF-8&cat=Computers

%2FSoftware%2FDatabases%2FData_Mining
%2FPublic_Domain_Software

4.3 Statistical methods in linguistic research

Mathematical and statistical concepts and techniques have long been used across a range of disciplines concerned in some sense with natural language, and these concepts and techniques are often relevant to corpus-based linguistics. Two such disciplines have just been mentioned: information retrieval and data mining. Others are natural language processing (Manning & Schütze 1999; Dale et al. 2000; Jurafsky & Martin 2008; Cole et al. 2010; Indurkhya & Damerau 2010), computational linguistics (Mitkov 2003), artificial intelligence (Russell & Norvig 2009), and the range of subdisciplines that comprise cognitive science (including theoretical linguistics) (Wilson & Keil 2001). The literatures for these are, once again, very extensive, and, to keep the range of reference within reasonable bounds, two constraints are self-imposed: (i) attention is restricted to the use of statistical methods in the analysis of natural language corpora for scientific as opposed to technological purposes, and (ii) only a small and, one hopes, representative selection of mainly through not exclusively recent work from 1995 onwards is given, relying on it as well as (Köhler & Hoffmann 1995) to provide references to earlier work.

Textbooks:

(Woods et al. 1986; Souter & Atwell 1993; Stubbs 1996; Young 1997; Biber et al.1998; Oakes 1998; Baayen 2008, Johnson 2008; Gries 2009, Gries et al. 2009).

Specific applications

As with other areas of science, most of the research literature on specific applications of quantitative and more specifically statistical methods to corpus analysis is in journals. The one most focused on such applications is the *Journal of Quantitative Linguistics*; other important ones, in no particular order, are *Computational Linguistics*, *Corpus Linguistics and Linguistic Theory*, *International Journal of Corpus Linguistics*, *Literary and Linguistic Computing*, and *Computer Speech and Language*.

- Language classification: (Cooper 2008; Kita 1999; Silnitsky 2003)
- Lexis: (Allegrini *et al.* 2000; Andreev 1997; Baayen 2001; Best 2001; Lin 1998; Lin & Ave 1998; Lin & Pantel 2001; Oakes & Farrow 2007; Romanov 2003; Yarowski 2000; Watters 2002)
- Syntax: (Gamallo et al. 2005; Gries 2001; Köhler & Altmann 2000; Köhler & Naumann 2007)
- Variation: (Cichocki 2006; Gooskens 2006; Heeringa & Nerbonne 2001, 2012; Hyvönen et al. 2007; Kessler 1995; Kleiweg et al. 2004; Nerbonne 2008, 2009, 2010; Nerbonne & Heeringa 2001; Nerbonne & Kretzschmar 2003; Nerbonne et al. 2008; Wieling & Nerbonne 2010; Wieling et al. 2011)

- Phonetics / phonology / morphology: (Andersen 2001; Cortina-Borja et al. 2002; Clopper & Paolillo 2006; Calderone 2009; Hubey 1999; Jassem & Lobacz 1995; Kageura 1999; Mukherjee et al. 2009; Sanders & Chin 2009)
- Sociolinguistics: (Macaulay 2009; Moisl & Jones 2005; Moisl et al. 2006; Moisl & Maguire 2008; Paolillo 2002; Tagliamonte 2006)
- Document clustering and classification: (Lebart & Rajman 2000; Manning & Schütze 1999; Merkl 2000). Document clustering is prominent in information retrieval and data mining, for which see the references to these given above.

Many of the authors cited here have additional related publications, for which see their websites and the various online academic publication directories.

Corpus linguistics websites

- Gateway to Corpus Linguistics: http://www.corpus-linguistics.com/
- Bookmarks for Corpus-Based Linguistics: http://personal.cityu.edu.hk/~davidlee/devotedtocorpora/CBLLinks.htm
- Statistical natural language processing and corpus-based computational linguistics: an annotated list of resources: http://nlp.stanford.edu/links/statnlp.html
- Intute. Corpus Linguistics: http://www.intute.ac.uk/cgi-bin/browse.pl?id=200492
- Stefan Gries' home page links: http://www.linguistics.ucsb.edu/faculty/stgries/other/links.html

- Text Corpora and Corpus Linguistics: http://www.athel.com/corpus.html
- UCREL: http://ucrel.lancs.ac.uk/
- ELSNET: http://www.elsnet.org/
- ELRA: http://www.elra.info/
- Data-intensive Linguistics (online textbook): http://www.ling.ohio-state.edu/~cbrew/2005/spring/684.02/notes/dilbook.pdf

REFERENCES

Allegrini, Paolo, Montemagni, Simonetta, Pirrelli, Vito (2000) 'Learning word clusters from data types'*, COLING-00*, 8-14.

Allen, Will, Beal, Joan, Corrigan, Karen, Maguire, Warren, and Moisl, Hermann 2007. ‘A linguistic time-capsule: The Newcastle Electronic Corpus of Tyneside English’, in J. Beal, K. Corrigan, H. Moisl (eds.) *Creating and Digitising Language Corpora, Vol. 2: Diachronic Databases.* Houndmills: Palgrave Macmillan, 16-48.

Andersen, Simone (2001) 'The Emergence of Meaning: Generating Symbols from Random Sounds - A Factor Analytic Model', *Journal of Quantitative Linguistics* 8: 101-136.

Andreev, S. (1997) 'Classification of Verbal Characteristics Based on their Implication Force by Means of Cluster- and Factor Analysis', *Journal of Quantitative Linguistics* 4: 23-29.

Arabie, P. Hubert, L., De Soete, G. (1996) *Clustering and Classification*, Singapore: World Scientific.

Baayen, R.Harald. (2001), *Word Frequency Distributions*. Dordrecht: Kluwer.

Baayen, R.Harald. (2008) *Analyzing Linguistic Data. A practical Introduction to Statistics using R*, Cambridge: Cambridge University Press.

Baker, Paul (ed.) (2009) *Contemporary Corpus Linguistics*. Continuum.

Best, Karl-Heinz (2001) 'Probability Distributions of Language Entities', *Journal of Quantitative Linguistics* 8: 1-11.

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

Calderone, Basilio (2009) 'Learning Phonological Categories by Independent Component Analysis', *Journal of Quantitative Linguistics* 16: 132-56.

Casella, George, Berger, Roger (2001*) Statistical Inference*, 2nd ed., Pacific Grove CA: Duxbury.

Chalmers, Alan 1999. *What is this thing called science?* 3rd ed. New York: McGraw-Hill / Open University Press.

Cichocki, Wladyslaw (2006) 'Geographic Variation in Acadian French /t/: What can Correspondence Analysis contribute toward Explanation?', *Literary and Linguistic Computing*, 21: 529-541.

Clackson, James (2007) *Indo-European Linguistics: An Introduction*, Cambridge: Cambridge University Press.

Clopper, Cynthia, Paolillo, John (2006) 'North American English Vowels: A Factor-analytic Perspective'. *Literary and Linguistic Computing*, 21: 445-462.

Cole, Ronald, Mariani, Joseph, Uszkoreit, Hans, Battista, Giovanni, Zaenen, Annie, Zampoli, Antonio (2010) *Survey of the State of the Art in Human Language Technology*, Cambridge: Cambridge University Press.

Cooper, Martin (2008) 'Measuring the Semantic Distance between Languages from a Statistical Analysis of Bilingual Dictionaries', *Journal of Quantitative Linguistics* 15: 1-33.

Cortina-Borja, Mario, Stuart-Smith, Jane, Valinas-Coalla, Leopoldo (2002) 'Multivariate Classification Methods for Lexical and Phonological Dissimilarities and their Application to the Uto-Aztecan Family', *Journal of Quantitative Linguistics*, 9: 87-124.

Dale, Robert, Moisl, Hermann, Somers, Harold. (2000) *Handbook of Natural Language Processing*, New York: Marcel Dekker.

Devore, Jay, Peck, Roxy (2005) *Statistics. The Exploration and Analysis of Data*, 5th ed., Florence KY: Thomson Brooks/Cole.

Everitt, Brian, Landau, Sabine, Leese, Morven, Stahl, Daniel (2011) *Cluster Analysis*, 5th ed., London: Arnold.

Fraleigh, John, Beauregard, Raymond (1994) *Linear Algebra*, 3rd ed., London: Addison-Wesley.

Freedman, David, Pisani, Robert, Purves, Roger (2007) *Statistics*, 4th ed., London: W.W.Norton.

Freedman, David (2009) *Statistical Models: Theory and Practice*, 2nd ed., Cambridge: Cambridge University Press.

Gamallo, Pablo, Agustini, Alexandre, Lopes, Gabriel (2005) 'Clustering Syntactic Positions with Similar Semantic Requirements', *Computational Linguistics*, 31: 107-145.

Gan, Guojun, Ma, Chaoqun, Wu, Jianhong (2007) *Data Clustering: Theory, Algorithms, and Applications*, SIAM, Society for Industrial and Applied Mathematics.

Gooskens, Charlotte (2006) 'The Relative Contribution of Pronunciational, Lexical, and Prosodic Differences to the Perceived Distances between Norwegian Dialects'. *Literary and Linguistic Computing*, 21: 477-492.

Gordon, A. (1999) *Classification*, 2nd ed. London: Chapman & Hall.

Gravetter, Frederick, Wallnau, Larry, (2008) *Statistics for the Behavioral Sciences*, 8th ed., Florence KY: Wadsworth.

Gries, Stefan. (2001), 'Multifactorial Analysis of Syntactic Variation: Particle Movement Revisited', *Journal of Quantitative Linguistics* 8: 33-50.

Gries, Stefan (2009) *Quantitative Corpus Linguistics with R*, New York: Routledge.

Gries, Stefan, Wulff, Stefanie, Davies, Mark (2009*) Corpus-linguistic Applications: Current Studies, New Directions*, Amsterdam: Rodopi.

Grimm, Laurence, Yarnold, Paul (2000) *Reading and understanding more multivariate statistics*, American Psychological Association.

Härdle, Wolfgang, Simar, Leopold (2007) *Applied Multivariate Statistical Analysis*, 2nd ed, New York: Springer.

Hand, David, Mannila, Heikki, Smyth, Padhraic (2001) *Principles of Data Mining*. Cambridge MA: MIT Press.

Hair, Joseph, Black, William, Babin, Barry, Anderson, Rolph (2007) *Multivariate Data Analysis*, 7th ed. New Jersey: Prentice-Hall.

Heeringa, Wilbert, Nerbonne, John (2001) 'Dialect areas and dialect continua', *Language Variation and Change* 13: 375-400.

Heeringa, Wilbert, Nerbonne, John (2012) Dialectometry. Accepted to appear in: Frans Hinskens & Johan Taeldeman (eds.): *Language and Space. An International Handbook of Linguistic Variation, Volume III: Dutch.* (Series: Handbook of Linguistics and Communication Science (HSK)). New York: Walter de Gruyter.

Hubey, H. (1999) 'Vector Phase Space for Speech Analysis via Dimensional Analysis', *Journal of Quantitative Linguistics* 6: 117-148.

Hyvönen, Saara, Leino, Antti, Salmenkivi, Marko (2007) 'Multivariate Analysis of Finnish Dialect Data—An Overview of Lexical Variation', *Literary and Linguistic Computing* 22: 271-290.

Indurkhya, Nitin, Damerau, Frederick (2010) *Handbook of Natural Language Processing*, 2nd ed., New York: Chapman & Hall.

Jain, Anil, Dubes, Richard (1988). *Algorithms for clustering data*, Englewood Cliffs, N.J.: Prentice Hall.

Jain, Anil Murty, M., Flynn, P. (1999) 'Data clustering: a review', *ACM Computing Surveys* 31: 264-323.

Jassem, Wiktor, Lobacz, Piotra (1995), 'Multidimensional Scaling and its Applications in a Perceptual Analysis of Polish Consonants', *Journal of Quantitative Linguistics* 2: 105-24.

Johnson, Keith (2008) *Quantitative Methods in Linguistics*, Oxford: Wiley-Blackwell.

Jurafsky, Daniel, Martin, James (2008) *Speech and Language Processing*, 2nd ed., Englewood Cliffs, N.J.: Prentice Hall.

Kachigan, Sam (1991*) Multivariate Statistical Analysis: A Conceptual Introduction*, 2nd ed., Nedw York: Radius Press.

Kageura, Kyo (1999) 'Bigram Statistics Revisited: A Comparative Examination of Some Statistical Measures in Morphological Analysis of Japanese Kanji Sequences', *Journal of Quantitative Linguistics* 6: 149-166.

Kaufman, Leonard and Rousseeuw, Peter 2005. *Finding Groups in Data. An Introduction to Cluster Analysis*. 2nd ed. Hoboken NJ: Wiley Blackwell.

Kennedy, Graeme (1998) *An Introduction to Corpus Linguistics*, New York: Addison Wesley Longman.

Kessler, Brett (1995)' Computational dialectology in Irish Gaelic'*, Seventh Conference of the European Chapter of the Association for Computational Linguistics*, University College Dublin, March 1995.

Kita, Kenji (1999) 'Automatic Clustering of Languages Based on Probabilistic Models', *Journal of Quantitative Linguistics* 6: 167-171.

Kleiweg, Peter, Nerbonne, John, Bosveld, Leonie (2004) 'Geographic projection of cluster composites', in: Blackwell, A. / Marriott, K. / Shimojima, A. (eds) *Diagrammatic Representation and Inference. Third International Conference, Diagrams 2004. Cambridge, UK, March 2004*. Berlin: Springer, 392-394.

Köhler, Reinhard, Altmann, Gabriel (2000) 'Probability Distributions of Syntactic Units and Properties' , *Journal of Quantitative Linguistics* 7: 189-200.

Köhler, Reinhard, Hoffmann, Christiane (1995) *Bibliography of quantitative linguistics*, Amsterdam: John Benjamins.

Köhler, Reinhard, Naumann, Sven (2007) 'Quantitative Text Analysis Using L-F- and T-Segments', *Data Analysis, Machine Learning and Applications, Proceedings of the 31st Annual Conference of the Gesellschaft für Klassifikation e.V.*, Albert-Ludwigs-Universität Freiburg, March 7–9, 2007, 637-645.

Lebart, Ludovic, Rajman, Martin (2000), 'Computing similarity', in: Dale *et al.* 2000, 477- 505.

Lin, Dekang (1998) 'Automatic retrieval and clustering of similar words', *COLING-ACL'98*, Montreal, 768-74.

Li, Hang, Abe, Naoke (1998) 'Word clustering and disambiguation based on co-occurrence data', *COLING-ACL'98*, Montreal, 749-55.

Lin, Dekang, Pantel, Patrick (2001) 'Induction of semantic classes from natural language text', *SIGKDD-01*, San Francisico.

Macaulay, Ronald (2009) *Quantitative Methods in Sociolinguistics*, Houndmills UK: Palgrave Macmillan.

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

Manning, Christopher, Raghavan, Prabhakar, Schütze, Hinrich (2008) *Introduction to Information Retrieval*, Cambridge: Cambridge University Press.

McEnery, Anthony, Wilson, Andrew (2001) *Corpus Linguistics*, 2nd ed., Edinburgh: Edinburgh University Press.

Merkl, Dieter (2000) 'Text data mining', in Dale *et al.* 2000, 889-903.

Mirkin, Boris (2005*) Clustering for Data Mining*, London: Chapman & Hall.

Mitkov, Ruslan 2005. *The Oxford Handbook of Computational Linguistics*. Oxford: Oxford University Press.

Moisl, Hermann 2007. 'Data nonlinearity in exploratory multivariate analysis of language corpora', in *Computing and Historical Phonology. Proceedings of the Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology*, *June 28 2007*, ed. Nerbonne, John, Ellison, Mark, and Kondrak, Grzegorz, Association for Computational Linguistics, 93-100. Available online at: http://www.let.rug.nl/alfa/Prague/proceedings.pdf.

Moisl, Hermann (2008) 'Using electronic corpora to study language variation: the problem of data sparsity', in Tsiplakou, S; Karyolemu, M; Pavlou, P, (eds.) *Language Variation. European Perspectives,* Amsterdam:John Benjamins, 169-178.

Moisl, Hermann (2009) ‘Sura length and lexical probability estimation in cluster analysis of the Qur'an’, *Association for Computing Machinery Transactions on Asian Language Information Processing 8.* Available online at: http://portal.acm.org.

Moisl, Hermann, Jones, Val (2005), 'Cluster analysis of the Newcastle Electronic Corpus of Tyneside English: a comparison of methods', *Literary and Linguistic Computing* 20: 125-46.

Moisl, Hermann, Maguire, Warren, Allen, Will (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: Meertens Institute.

Moisl, Hermann, Maguire, Warren (2008). 'Identifying the Main Determinants of Phonetic Variation in the Newcastle Electronic Corpus of Tyneside English', *Journal of Quantitative Linguistics* 15: 46-69.

Myatt, Glenn (2006) Making *Sense of Data: A Practical Guide to Exploratory Data Analysis and Data Mining,* Hoboken NJ: Wiley-Interscience.

Myatt, Glenn, Johnson, Wayne (2009) *Making Sense of Data II: A Practical Guide to Data Visualization, Advanced Data Mining Methods, and Applications*, Hoboken NJ: Wiley.

Mukherjee, Animesh, Choudhury, Monojit, Basu, Anupam, Ganguly, Niloy (2009) 'Self-organization of the Sound Inventories: Analysis and Synthesis of the Occurrence and Co-occurrence Networks of Consonants', *Journal of Quantitative Linguistics* 16: 157-184.

Nerbonne, John (2008) ‘Variation in the aggregate: an alternative perspective for variationist linguistics’. In: Kees Dekker, Alasdair MacDonald and Hermann Niebaum (eds.), *Northern Voices: Essays on Old Germanic and Related Topics offered to Professor Tette Hofstra,* 365–382*.* Leuven: Peeters.

Nerbonne, John (2009) 'Data-driven Dialectology', *Language and Linguistics Compass* 3: 175-198.

Nerbonne, John (2010) Mapping Aggregate Variation. In: Alfred Lameli, Ronald Kehrein and Stephan Rabanus (eds.) *Language and Space. International Handbook of Linguistic Variation. Vol. 2 Language Mapping*. Berlin: Mouton De Gruyter. 2010. Chap. 24. pp. 476-495, maps pp.2401-2406.

Nerbonne John, Heeringa Wilbert (2001), 'Computational comparison and classification of dialects', *Dialectologia et Geolinguistica* 9: 69-83.

Nerbonne, John, Kretzschmar, William (2003) 'Introducing computational methods in dialectometry*', Computers and the Humanities* 37: 3245-3255.

Nerbonne, John, Kleiweg, Peter, Heeringa, Wilbert, Manni, Franz (2008) ‘Projecting Dialect Differences to Geography: Bootstrap Clustering vs. Noisy Clustering’. In: Christine Preisach, Lars Schmidt-Thieme, Hans Burkhardt & Reinhold Decker (eds.) *Data Analysis, Machine Learning, and Applications. Proc. of the 31st Annual Meeting of the German Classification Society* Berlin: Springer, 647-654.

Nisbet, Robert, Elder, John (2009) *Handbook of Statistical Analysis and Data Mining Applications*, New York: Academic Press.

Oakes, Michael (1998) *Statistics for Corpus Linguistics*, Edinburgh: Edinburgh University Press.

Oakes, Michael, Farrow, Malcolm (2007) 'Use of the Chi-Squared Test to Examine Vocabulary Differences in English Language Corpora Representing Seven Different Countries', *Literary and Linguistic Computing* 22: 85-99.

Paolillo, John (2002) *Analyzing Linguistic Variation. Statistical Models and Methods*, Stanford: CSLI Publications.

Popper, Karl (1959) *The Logic of Scientific Discovery*. New York: Basic Books.

Popper, Karl (1963) *Conjectures and Refutations: The Growth of Scientific Knowledge*. Florence KY: Routledge / Taylor & Francis Group.

Poole, David (2005) *Linear Algebra: A Modern Introduction*. Florence KY: Brooks Cole.

Rice, John (2006) *Mathematical Statistics and Data Analysis*, 3rd ed., Duxbury.

Romanov, Yuri (2003) 'Investigation of English Verb Properties by Means of the Correlation Analysis Method', *Journal of Quantitative Linguistics* 10: 117-127.

Romesburg, Charles (1984) *Cluster Analysis for Researchers*. Florence KY: Wadsworth.

Russell, Stuart, Norvig, Peter (2009) *Artificial Intelligence: A Modern Approach*, 3rd ed, New Jersey: Prentice Hall.

Sanders, Nathan, Chin, Steven (2009) 'Phonological distance measures', *Journal of Quantitative Linguistics* 16: 96-114.

Silnitsky, George (2003) 'Correlation of Phonetic and Morphological Systems of Indo-European Languages', *Journal of Quantitative Linguistics* 10: 129-141.

Souter, Clive and Eric Atwell (1993), *Corpus-Based Computational Linguistics.* Amsterdam: Rodopi.

Strang, Gilbert (2009) *Introduction to Linear Algebra*, 4th ed., Wellesley: Cambridge Press.

Stubbs, Michael (1996) *Text and Corpus Analysis*: *Computer-assisted Studies of Language and Culture*, Oxford*:* Wiley-Blackwell*.*

Tagliamonte, Sali (2006) *Analysing Sociolinguistic Variation*, Cambridge: Cambridge: Cambridge University Press.

Watters, Paul (2002) 'Discriminating English Word Senses using Cluster Analysis', *Journal of Quantitative Linguistics* 9: 77-86.

Wieling, Martijn, Nerbonne, John (2010) ‘Bipartite spectral graph partitioning for clustering dialect varieties and detecting their linguistic features’. *Computer Speech and Language* 25: 700-715.

Wieling, Martijn, Shackleton, Robert, Nerbonne, John (2011) ‘Analyzing Phonetic Variation in the Traditional English Dialects: Simultaneously Clustering Dialect and Phonetic Features’. Submitted to *LLC: Journal of Digital Scholarship in the Humanities*.

Wilson, Robert, Keil, Frank (2001) *The MIT Encyclopedia of the Cognitive Sciences*, Cambridge MA: MIT Press.

Woods, Anthony, Fletcher, Paul, Hughes, Arthur (1986) *Statistics in Language Studies*. Cambridge: Cambridge University Press.

Xu. Rui, Wunsch, Don (2008) *Clustering*, London: Wiley-IEEE Press.

Yarowsky, David (2000) 'Word-sense disambiguation', in Dale *et al.* 2000, 629-654.

Young, Steve, Bloothooft, Gerrit (1997) *Corpus-based Methods in Language and Speech Processing*, Dordrecht: Kluwer Academic.