CLUSTER ANALYSIS OF LEGAL DOCUMENTS

by

JILLIAN BOREHAM

A Thesis Submitted for the Degree of

Doctor of Philosophy

at the

UNIVERSITY OF KENT AT CANTERBURY

Canterbury August, 1976.

CONTAINS

PULLOUTS

ABSTRACT

Single-link cluster analysis has been used to provide classifications of several collections of legal documents, based on various characteristics of the text. Each document was represented in terms of the chosen characteristics by a vector whose elements were the frequencies of occurrence of the characteristics in that document. The values of similarity between documents were determined by calculating the cosine of the angle between each pair of document vectors. The clustering algorithm then operated on these similarity coefficients to group documents which were most similar.

A suite of computer programs was written to perform the classification. Four programs were required to (a) select the document descriptors from the full-text of the documents, (b) construct document vectors, (c) calculate similarity coefficients, and (d) perform single-link clustering.

Three classification experiments were performed. The first classified the full-text of both the English and French versions of the Treaties of the Council of Europe. The words of the full-text, taken singly and in pairs, were used to describe the treaties, and the two cases of including and excluding the 'common' words were investigated. The best classification was based on single words with common words excluded. Since each treaty was a lengthy collection of non-homogeneous clauses, it was thought that a classification

of the individual articles would be more useful. In this case the formal and non-formal clauses clustered separately, whereas before the formal clauses, present in every treaty, had caused semantically unrelated treaties to be brought together.

During the course of this study an opportunity arose to investigate the use of cluster analysis to test the trustworthiness of certain oral confessions presented as evidence in criminal proceedings. The common or function words, which are generally agreed to characterise the style of an author, were used as document descriptors for two sets of statements, one which the defendant admitted, the other which he was alleged to have made but which he denied. The two sets of statements clustered separately, indicating a difference in style. On the basis of this and other comparative tests it was possible to say that the disputed statements were unlikely to have been made by the defendant.

The third experiment involved the use of the marginal citations in Statutes as document descriptors. Statutes were regarded as semantically related if they cited the same Acts. The Public General Acts of Parliament for the three years 1973 - 1975 were successfully clustered into groups of related Acts.

The vector representation can also be used as a basis for a document retrieval system. An experimental vector search system was developed as part of this study. A question is input in natural language, converted to vector form and compared with the document vectors to determine the degree of similarity. The result of the search consists of a list of the relevant documents, ranked in decreasing order of similarity with the question. The system performed encouragingly during small-scale tests, and this thesis discusses possible future developments of such a system.

ACKNOWLEDGEMENTS

The author wishes to thank the Atomic Energy Research Establishment, Harwell, for the financial support provided throughout the three years of this work.

The experimental work described in Chapter VIII was carried out in collaboration with Mrs. Margaret Anderson, and her help is gratefully acknowledged.

Finally, the author thanks her supervisor, Dr. Bryan Niblett, for his guidance and enthusiasm throughout this project.

CONTENTS

Contents

Contents

CHAPTER I

INTRODUCTION

I.1 Automatic text processing

> "The attraction of law texts as a data base is that the language is, to a degree, already formalized: . . . (the) texts are drafted by lawyers who spend a considerable effort deciding the choice of words and sentence constructions, aiming to convey meanings and intentions unambiguously."

Thus states Myers (1) in his review of computerised legal data processing.

In this thesis the formal nature of legal documents has been exploited in order to study the classification of such documents by computer. A full-text method is used, based on the occurrences of words in the text. This is in contrast to most automatic methods which have been developed in the past for classifying scientific documents, and which operate on sets of index terms assigned to documents.

A full-text or natural language system is one in which each document is represented by the complete set of words which occur in the text of the document. The alternative is the indexed system where documents are described by index terms or keywords which summarise the contents. Usually these terms are controlled so that only a restricted, prescribed set of terms are available for describing the documents.

One of the main arguments against the full-text method is that authors are inconsistent in their use of vocabulary. Different authors use different words to describe the same concepts, and often the same author will vary his expression of a particular idea in order to make his writing more interesting. However these problems are less evident in legal documents because of the formal vocabulary and structure which results in more consistency amongst legal authors. Moreover, as Tapper (2) and Borko (3) point out, indexers are no more consistent in assigning index terms to documents.

Another argument against indexing is that the indexer is not likely to be as expert in the specialised subject of a document as its author. The indexer is unlikely to improve on the author's choice of words.

A significant advantage of the full-text method is that it does not require extensive preparation of the data for input to the system. Indexing is both costly and time-consuming, and therefore best avoided unless some really worthwhile advantage is to be gained from it. The Aberystwyth indexing language experiment, described by Keen (4), concluded that uncontrolled full-text systems offer the best all-round performance, in terms of ease of use and search results.

Moreover, many document collections are now readily available in machine readable form as a by-product of computer typesetting operations, and it is convenient to use this data without further preparation.

I.2 Aims of document classification

Classification is fundamental to most means of communication between people. For example, words are used to describe classes of things and ideas having similar properties. This thesis is itself classified into chapters, each dealing with a homogeneous group of ideas, and this classification is desirable for a clear understanding of the ideas presented. Subjectively it is obvious what criteria have been used for choosing the chapters, but it is not easy to formalise these.

Libraries have been using classification schemes for many years to organise large quantities of information in document form. The most widely used of the conventional library classifications are those of Dewey, dating back to 1876, the Library of Congress devised in 1901 and the Universal Decimal Classification, 1905. For details of these and other library classifications see Sayers (5).

These library schemes attempt to subdivide the whole 'Universe of knowledge' into logical classes to which documents are assigned in the most appropriate way.

As the various fields of knowledge grow and develop more specialised areas this system of classification breaks down because it is impossible to specify beforehand all the classes which will be required to accommodate the diversity of documents produced. In general, an information seeker is interested in the details of one particular subject, and it is helpful to him if this subject has its own 'tailor-made' classification scheme which can be regularly updated to incorporate new ideas as they arise.

For a flexible classification which is easy to update we need to use an a posteriori method in contrast to the a priori methods of conventional classification. We require a knowledge of the contents of all the documents to be classified beforehand; these can then be grouped according to similarity of content. A large number of ideas must be associated simultaneously as no one idea in a document can be singled out to characterise that document completely. This is where the a priori methods fail, by specifying insufficient classes to accommodate the various combinations of ideas that can occur. Sokal (6), however, has pointed out the difficulty of associating large numbers of ideas.

> "It is obviously much more complicated to establish classifications based on many characters than it is to establish classifications on only one character. The human mind finds it difficult to tabulate and process large numbers of characters without favouring one aspect or another."

To overcome this problem numerical, objective techniques have been developed. These are discussed in greater detail in Chapter III. We are concerned here with the uses of a classification scheme rather than the details of its construction. We deal specifically with the computerised implementation of numerical classification methods, which is essential for large document collections because of the vast amount of computation involved. The following section briefly discusses some of the experiments in computerised document classification.

I.3 Automatic classification experiments

Because there are, as yet, very few useful linguistic aids available for automatically analysing text, most automatic classification schemes are based on the statistics of word occurrences in documents. We assume that the words used in a document serve as an indication of the subject class to which that document belongs, though this is not an ideal basis for classification as Maron (7) points out:

> "It is because words and sentences stand for other things (i.e., they are one step removed from the things and events that they describe) that the problem of indexing information is made even more complex than the problem of classifying non linguistic entities. More specifically, the problem is so complex not only because words and sentences are one step removed, but also because there is no one-to-one relationship between the individual words and the events that the sentence containing those words describes."

Despite this fundamental difficulty, reasonable results have been obtained using the occurrences of words as a basis. Maron's own work combines automatic and a priori methods. A fixed number of classes are specified beforehand, and documents are assigned to these classes automatically according to the value of a probabilistic measure based on the presence or absence in the documents of preselected keywords.

An alternative method by Borko and Bernick, described in references (8) and (9), uses factor analysis to classify documents. In this case the classes are not set up in advance, but are derived automatically from the documents.

A system which automatically classifies documents and uses the resulting classification as an aid for information retrieval is decribed by Hoyle (10). Such uses of classification schemes are discussed in section I.4.

The SMART system devised by Salton includes automatic classification schemes and reference (11) contains several reports on this aspect of the system. Included are Dattola's work (12) on fast methods of automatic classification which are feasible for very large collections, and Worona's description (13) of a method of clustering questions. We discuss the use of question clustering as a search aid in section I.4.1.

A further experiment by Salton (14) classifies documents, not by their word content, but on the basis of the bibliographic citations they contain. This can be done in two ways:

(a) by grouping together documents which are cited by the same documents;

or (b) by grouping together documents which cite the same documents.

For early documents on a particular topic, method (a) is the better because these documents are likely to have been cited many times, enabling associations to be formed. However they will not themselves cite many documents as there will be little earlier literature on the subject. Conversely, method (b) is preferable for later documents produced after the subject has become well-established. There will be a wealth of literature for these documents to cite, but obviously they will not have been frequently cited themselves.

Sparck Jones and Jackson (15) use the method of clumping to classify documents. This is an iterative procedure which begins by defining a 'cohesion function' on the values of some measure of association between documents, and attempts to maximise this function on groups of documents.

The method used in this study for classifying legal documents is that of single-link clustering, which is described in detail in Chapter III. In reference (16) van Rijsbergen uses this method and describes how to search the resulting arrangement of classified documents.

For a fuller summary of experiments in automatic classification see the table presented by Prywes and Smith (17), which includes details of the text used, the number of words in the documents and the number of keywords assigned, the classification algorithm and the size and number of classes in the resulting scheme.

I.4 Uses of classification schemes

Classification, in general, can be regarded as a compromise of two conflicting processes:

(i) an information-gaining process - information about two documents is gained from the fact that they belong to the same class. Their relationship provides more information about both of them than if each were considered independently.

(ii) an information-losing process - the detailed characteristics of individual documents are lost when only the classes to which they belong are considered. The classes themselves are characterised by just those properties which are used for classificatory purposes, and which are responsible for bringing the members of a particular class together. Other properties of individuals which do not contribute to the classification are ignored.

We classify documents in order to exploit the information gained in the process; that is, to use the relations between documents to gain a clearer understanding of their contents. In the following sections we discuss the use of classification schemes in information retrieval systems.

I.4.1 Search tactics

Classification schemes can be incorporated into retrieval systems to improve the performance measured in terms of Recall and Precision. These measures are usually defined as follows:

$$\text{Recall} = \frac{\text{number of relevant documents that are retrieved}}{\text{total number of relevant documents in the collection}}$$

$$\text{Precision} = \frac{\text{number of retrieved documents that are relevant}}{\text{total number of documents retrieved}}$$

Associative retrieval methods benefit most readily from classification schemes. These methods operate by calculating the value of some function which measures the association between the question and each document in the collection, in terms of the words or keywords they contain. (Measures of association are discussed in detail in III.5) The values of the function indicate the degree of relevance of each document to the question asked, and the set of documents to be retrieved is decided by these values.

The simplest implementation of this technique is Linear Associative Retrieval, which ranks the documents in order of the values of the association function, and retrieves those documents lying above some specified cut-off point. This involves lengthy comparisons of the question with every document in the collection. However, the number of comparisons can be reduced considerably by incorporating a classification scheme. First we generate a classification of all the documents to be searched, and for each class choose some representative to be used in place of the individual documents. The question is initially compared with each of the class representatives, to calculate the corresponding values of the association measure. Those representatives yielding a high value are chosen for the next stage. The linear associative method is then applied to the documents contained in the classes whose representatives are chosen.

Salton (18) has shown that the SMART system using this form of cluster searching performs better than the MEDLARS system using a traditional Boolean strategy. The success of the method depends on the choice of the class representatives. A bad representative can cause a whole class of relevant documents to be rejected, since it may have itself a low value of association with the question. This demonstrates the information-losing aspect of classification; we miss relevant documents by considering classes rather than individuals.

A more sophisticated way of using classes is to arrange them into a hierarchy, that is a tree structure of nested classes. The bottom of the hierarchy consists of the individual documents, each in a class by itself. At the top the classification is complete, and we can consider the root of the tree to be the single class containing all the documents.

There are two ways of searching a hierarchy, either (i) from the top down, or (ii) from the bottom up.

(i) Beginning at the top of the hierarchy we compare the question with each of the cluster representatives and choose the best matching one. At the next stage the question is compared with the representatives of all the sub-clusters contained in the cluster chosen at the previous stage. We work down the hierarchy, repeating this process until an optimal representative is found; that is, when the best matching representative at the next stage down produces a worse match than the best one at the present stage, which becomes the optimal representative. The result of the search is the set of documents contained in the cluster represented by the optimal representative.

This process can be made more complex by selecting more than one cluster representative at each stage, if several produce good enough matches with the question. Eventually several optimal representatives will be found, and documents retrieved from the classes these represent. Recall is likely to be improved in this way, as all the relevant documents are unlikely to belong to a single cluster.

(ii) Alternatively, we may begin at the bottom of the hierarchy with the individual documents. We need to know beforehand of one relevant document in the collection, and proceed to work up the hierarchy to the largest cluster containing the given document which does not exceed some given threshold. This threshold may either set a limit on the size of the cluster, or specify a limiting value for the measure of association between the question and cluster representative.

The threshold, or cut-off point, can be used to optimise either recall or precision. A low cut-off value, meaning either that the maximum number of documents to be retrieved is small, or that the search is to be terminated at a low level in the hierarchy, will result in high precision but low recall. Conversely, a high cut-off value will give high recall and low precision.

The construction of a hierarchy does of course involve a large initial investment of computer time. However, averaged out over all searches which use the hierarchy this does not significantly increase the computer effort required for each individual search. Moreover, the effort in constructing the hierarchy is compensated by the increased speed of searching, since it is unnecessary to examine every document. Jardine and van Rijsbergen (19) have shown that an ideal hierarchical cluster search is more effective than an ideal linear associative search.

Question classification can also improve the efficiency of searching by using the results of previous searches. A record is kept of all previous questions together with their resulting sets of retrieved documents. Each new question can then be classified with respect to these stored questions. For each of the old questions in the class to which the new question belongs, the corresponding retrieved document set is selected, and the new question compared with these documents to find the relevant ones. For very large collections this can reduce considerably the number of documents examined.

Of course all these methods which restrict the search to a subset of the document collection necessarily impose an upper limit on the recall, determined by the number of relevant documents in the restricted collection. There may be other relevant documents in the rest of the collection which are unobtainable, however good the matching strategy.

We may, on the other hand, use a classification scheme to improve recall. We can expand the results of a conventional search, which has access to the whole collection, by retrieving all the documents in the cluster to which the initially retrieved documents belong. A very specific question can be asked in order to retrieve the highly relevant documents, and then expanding the results as above will provide documents which are relevant at a more general level.

For classification to be a successful search aid, we require document classes to correspond closely to relevance classes, that is, the sets of documents which are relevant to particular questions. Ideally we need our classification method to use the same processes as searching, and for associative searching we use the same measure of association for generating document classes and for comparing questions with documents.

I.4.2 Efficient storage of data

The retrieval time for any search strategy is proportional to the time taken to search the storage medium for the records relating to the relevant documents. This consists of the time taken to find the first relevant record plus the time taken to move from this record to the next and subsequent relevant records. In practice, the time needed to read a record is negligible compared to this 'memory access' time. If all the relevant records relating to a

particular question are stored in neighbouring locations, then only one costly memory access is required.

Obviously it is not possible to say beforehand which groups of documents will be relevant to questions asked in the future. However a classification of the documents can give some indication of this, on the assumption that closely associated documents will be relevant to the same question. Thus members of the same class can be stored in neighbouring locations, and provided the relevant documents for a particular question are spread over a small number of classes, the memory accesses will be kept to a minimum.

I.4.3 Browsing

Another useful application of a classification scheme is in browsing. Often we have only a vague idea of what information we are seeking, and need to browse through the document collection until we find something suitable or are able to formulate our need in terms of a formal question to the system. It is not particularly helpful to have to wade through the whole collection in a random manner, and a classification scheme, especially a hierarchical one, can guide us to the relevant sections. If we have a very general idea to begin with, we can find the class which satisfies this, and then browse down the hierarchy to find more specific documents. Alternatively, if we know of one document of interest, we may browse upwards to see what other documents are related to it at both specific and general levels.

We are thus able to browse through the document collection without being familiar with the indexing and search languages, and without having to formulate a question. Browsing in this manner is most effective when implemented on-line to a computer, though an automatic classification scheme can also be used in printed form.

The ideas for browsing presented here were suggested by Prywes and Litofsky (20) who discuss them in more detail.

I.5 Value of automatic classification

Early attempts at automatic classification were criticised for being inconsistent with human classification. However, as Swanson (21) points out:

> ". . . even though machines may never enjoy more than a partial success in library indexing, . . . people are even less promising."

That is to say, comparison with manual systems is inappropriate for evaluating an automatic classification, since manual classifications are themselves inconsistent. The fact that an automatic scheme disagrees with a manual one may mean that the automatic one is superior, though critics have usually taken it to mean the reverse.

In its favour, an automatic method will give a less biased classification since it is based on purely objective measures of association between documents.

> "The mere fact that the human classifier observes his classification as he constructs it, so that he necessarily finds it plausible, may mean that he does not examine the principles on which it is based in a sufficiently critical way, or even that he does not formulate them properly or apply them consistently."
>
> Sparck Jones (22).

Automatic classification is best evaluated in terms of its performance as an aid to retrieval, or for whatever other function it is constructed. This aspect of classification is discussed further in Chapter III.

I.6 Conclusion

The preceding sections have introduced automatic processing of the full text of documents, and have described the aims and uses of classification. The remainder of the thesis investigates the automatic classification of legal documents using the full text, in order to examine how well the words used in the documents represent the semantic content.

The classification method used is that of cluster analysis, an automatic numerical technique. This entails describing the document collection by a set of attributes, in this case the words in the documents, and representing each document in terms of these attributes by assigning appropriate values to them. Clustering is achieved by

comparing these document descriptions and grouping together those which are similar. To simplify the comparison of document descriptions, these are arranged to form vectors. Chapter II discusses the mathematical theory of vectors and how this applies to the vector representation of documents. Chapter III deals with the theory of cluster analysis, and describes various strategies for clustering, with special emphasis on the single-link method used in this study.

The computer programs used to implement the classification method are described in Chapter IV. These programs have been used to classify a document collection consisting of the Conventions and Agreements of the Council of Europe, and the results are presented and discussed in Chapter V. Both the English and French versions of these documents were used, and some interesting statistics of this data, obtained as a by-product of the classification process, are given in Chapter VI. Comparisons are made between the vocabularies of the two languages English and French as used in this particular document collection.

Chapter VII describes a novel application of cluster analysis to test the authenticity of certain oral confessions used as evidence in criminal proceedings.

In Chapter VIII a further application of cluster analysis is considered. The marginal citations are extracted from Statutes and their frequencies of occurrence are used as vector elements. Documents are then clustered together

if they cite the same Acts. This is analagous to method (b) described in section I.3 in connection with Salton's technique for classifying documents on the basis of their bibliographic citations.

Chapter IX considers a search strategy based on the classification techniques developed in earlier chapters. Searching is viewed as a form of classifying documents on the basis of their relevance to the question asked. The characteristics of the search method are compared with those of a Boolean method. The system is still at an early experimental stage, and further suggestions are given for developing it.

Finally, Chapter X summarises the important results of earlier chapters and discusses future developments of the methods investigated in this thesis.

CHAPTER II

VECTOR THEORY

II.1 Introduction

In this chapter we introduce the vector as a means of representing documents in terms of the frequencies of occurrence of the words they contain. The vector representation provides a convenient means of comparing documents to find how similar they are in word content.

The mathematical definitions of groups, fields and vector spaces are given in Appendix A. These definitions state the properties of vectors, which are the members of a vector space. An example of a vector space is the set of geometrical vectors in an n-dimensional space; that is, lines having both length and direction specified by sets of values for the n coordinates. We interpret document vectors as such lines in an n-dimensional space.

(Further mathematical discussion of vector spaces can be found in Patterson and Rutherford (23).)

II.2 Vector representation of documents

Consider a set of documents D. We may choose a set of n distinct attributes $\{A_1, A_2, A_3, \ldots A_n\}$ which represent the information content of the document collection.

This set of attributes constitutes an indexing vocabulary. To describe each individual document in terms of these attributes, we assign some value f_i to each attribute A_i, to indicate the extent to which the idea represented by A_i is included in the given document.

For one such document d contained in D, suppose the values of the n attributes are $f_1, f_2, f_3, \ldots f_n$. Then we may represent d by an n-tuple, an ordered set of n elements $(f_1, f_2, f_3, \ldots f_n)$, so that the ith element of the n-tuple is precisely the value assigned to A_i.

We call this n-tuple the 'document vector' for d and denote it by $\underline{d}$. The set of vectors for all the documents in D is denoted by $\underline{D}$.

II.2.1 Choice of attributes

As we wish to make use of the full-text of documents, the obvious choice for the attributes, or index terms, A_i is the set of n distinct words occurring in the document collection. Of course there are certain words such as prepositions, conjunctions and general terms which do not contribute much to the meaning of the text, but are merely functional. They are usually named 'common words', and are found to occur very frequently in most documents. These words are not particularly useful for describing documents, and may be excluded from the attribute list. This is the only form of control exerted over the indexing vocabulary derived from the full text.

It is convenient to refer to the word corresponding to attribute A_i as 'word i'.

II.2.2 Choice of attribute values

We now assign a value f_i to each word i to indicate its importance in a given document. This process is referred to as 'weighting' of index terms. A review of work in this subject is given by Sparck Jones (24) pp. 4.9 - 4.11.

The weighting function chosen for this study of classification is the simple document frequency. For a given document, the value assigned to each word i is the frequency of occurrence of that word in the document. In reference (24), Sparck Jones describes how, for searching, the document frequency performs less well than weighting functions which take into account the frequency of occurrence of each word in the whole collection. For example, the weighting applied to a word in a particular document might be given by the frequency of occurrence of the word in that document divided by the frequency of occurrence of the word in the whole collection. These collection frequency weighting functions give a high weighting to rare terms, which are useful for separating out the relevant documents in answer to a question. Further, words distributed evenly over a large number of documents are given lower weightings in order to avoid retrieving too many documents.

When classifying, however, we wish to identify the overlap of documents with respect to their word content. The very rare words are not so useful in this case, and the frequent words with a wider distribution demonstrate the relationships between documents.

A comment by Sparck Jones (25) sums up the essential difference between the two forms of weighting mentioned:

> "Weighting by collection frequency as opposed to document frequency is quite different. It places greater emphasis on the value of a term as a means of distinguishing one document from another than on its value as an indication of the content of the document itself."

It is precisely the contents of the documents that we are interested in, and how well these contents are described by the words used.

Two further points provide additional justification for the choice of document frequency weighting. It is convenient to compute these frequencies as each document is entered into the system. Collection frequencies, on the other hand, are not available until every document has been processed, and are usually computed at the search stage when used for retrieval, for which they are more suitable. Moreover, updating the document collection is more difficult when collection frequencies are involved. When a new document is added, the collection frequencies of those words contained in the new document alter, so that the vectors

for all existing documents containing these words must be regenerated. In contrast, the document frequencies do not change with the addition of new documents, and a vector has to be regenerated only when an amendment is made to the actual document it represents.

II.3 Interpretation of vector representation

We now have a vector representation for each document, in which the ith element is the frequency of occurrence f_i of word i in that document. These values f_i are all positive integers or zero; that is, they belong to the set of natural numbers. A value of zero for f_i means that word i does not occur in that document. In fact a large proportion of the f_i will be zero for any one document, because only a small number of the total n distinct words in the collection occur in any particular document.

We would like to interpret this document vector in terms of mathematical vector theory. In particular we wish to represent the vectors geometrically as described in section II.1; that is to show that each document vector lies in an n-dimensional vector space. In this case the n-tuple $(f_1, f_2, f_3, \ldots . f_n)$ determines a unique point in the space whose coordinate directions are given by the distinct words in the document collection. The straight line joining the origin to this point depicts the document vector, see Figure II.3 . (Note that only two of the coordinate directions can be shown.)

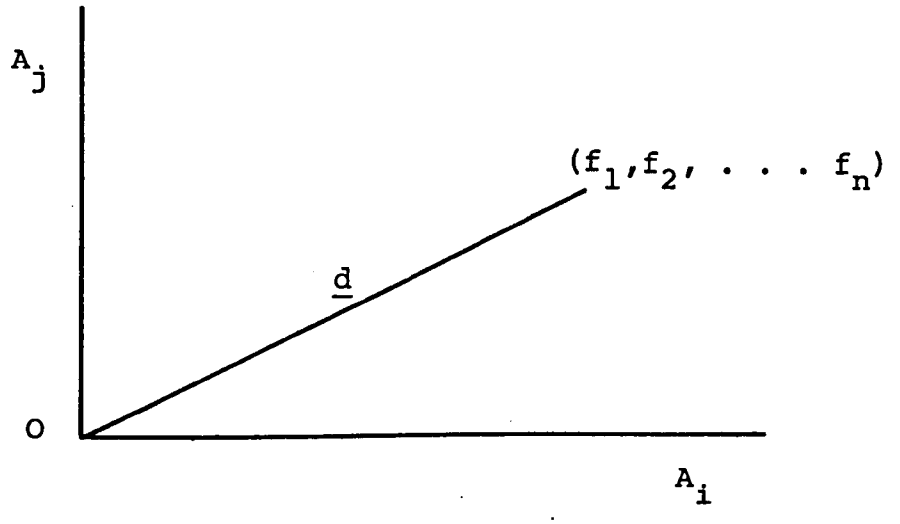


FIGURE II.3 Geometrical representation of document vector

We show in Chapter III that the angle between any two of these vectors is a measure of the degree of similarity between the two documents they represent.

In the following section we show how the mathematical properties of vectors can be applied to document vectors. We also show that the set of document vectors D does not itself constitute a vector space, but that we can expand D to produce a vector space of which D is a subset.

II.3.1 Document vector space

The conditions required for groups and vector spaces referred to in this section are given in the definitions in Appendix A.

We show the mathematical properties which the document vector possesses, and prove that $\underline{D}$ is not a vector space by showing that it is not even a group, hence it does not satisfy condition (1) for a vector space (see Appendix A).

We can define a binary operation on our set of vectors $\underline{D}$ as follows:

Let d and d' be any two documents in D, and let their vectors in $\underline{D}$ be given by

$$\underline{d} = (f_1, f_2, \ldots . f_n)$$

$$\underline{d}' = (g_1, g_2, \ldots . g_n)$$

Then define an operation + on $\underline{D}$ such that

$$\underline{d} + \underline{d}' = (f_1 + g_1, f_2 + g_2, \ldots . f_n + g_n)$$

where $f_i + g_i$ is the usual addition of two natural numbers, so that $f_i + g_i$ is another natural number.

The value $f_i + g_i$ is the frequency of occurrence of word i in the document d + d' which the vector $\underline{d} + \underline{d}'$ represents. One interpretation of $\underline{d} + \underline{d}'$ is that of a vector representing the concatenation of the texts of the two documents d and d'. We see presently, however, that this is not the only interpretation, nor is it the most appropriate.

Now the document d + d' may not belong to the collection D, but we can always extend the collection to include all possible d + d' and thus ensure that for every $\underline{d}$ and $\underline{d}'$ in $\underline{D}$ the vector $\underline{d} + \underline{d}'$ also belongs to $\underline{D}$; that is, the set $\underline{D}$ is closed under addition. The vector set may become infinite under this extension, but we will not have altered the coordinate system, as no new index terms will have been introduced.

We see already that the original set $\underline{D}$ might not be closed under addition and hence not a group (i.e. if it does not satisfy condition (1) for a group), but we will assume that it is closed, since we can always extend it so that it does satisfy this condition.

By the commutativity and associativity of the natural numbers under addition it follows that the operation + on $\underline{D}$ is commutative and associative.

For example,

$$\underline{d} + \underline{d}' = (f_1 + g_1, f_2 + g_2, \ldots . f_n + g_n)$$
$$= (g_1 + f_1, g_2 + f_2, \ldots . g_n + f_n)$$

by the commutativity of the natural numbers

$$= \underline{d}' + \underline{d}.$$

Similarly,

$$\underline{d} + (\underline{d}' + \underline{d}") = (\underline{d} + \underline{d}') + \underline{d}".$$

Hence addition on $\underline{D}$ satisfies conditions (2) and (3) for a group. Following our previous interpretation of addition, we see that associativity means the same document is produced by concatenating d and d' and appending d", as is produced by first concatenating d' and d" and appending this joint document to d. The resulting text is the same in both cases.

The commutativity property is not so obviously true because the order of words in a document is essential to the meaning they convey and commuting the documents re-arranges the text. A feature of the vector representation is that it destroys the order of the words in a document, and conveys information on word occurrence only. (This problem is dealt with further in section II.4) We see therefore that commutativity means merely that the document represented by the vector $\underline{d} + \underline{d}'$ contains the same words with the same frequencies of occurrence as the document represented by $\underline{d}' + \underline{d}$. It does not mean they are the same document, as the words may be arranged differently in each one.

From this it is clear that our first interpretation of $\underline{d} + \underline{d}'$ as representing the concatenation of the documents d and d' was too strict, and we can say only that $\underline{d} + \underline{d}'$ is the vector for the document containing the same words with the same frequencies as the combined contents of d and d'. Associativity says that this holds true for the addition of three documents, regardless of the order of the addition.

To satisfy condition (4) for a group we can define an identity element $\underline{\phi}$ in $\underline{D}$ by,

$$\underline{\phi} = (h_1, h_2, \ldots . h_n) \text{ where } h_i = 0, \text{ for } i=1, \ldots n.$$

$\underline{\phi}$ is the vector representing a document containing no words, a null document. It has the property that for all vectors in $\underline{D}$

$$\underline{d} + \underline{\phi} = \underline{d}$$

This follows from the fact that zero is the identity element for the natural numbers. The interpretation is that adding no words to a document results in a document containing the same words with the same frequencies as in the original, though the words may be re-arranged.

Finally we show that condition (5) for a group is <u>not</u> satisfied by $\underline{D}$ under addition. For a vector $\underline{d}$ in $\underline{D}$ to have an inverse, we need to find a vector $\underline{d}'$ in $\underline{D}$ such that

$$\underline{d} + \underline{d}' = \underline{\phi}$$

That is, we require a document whose contents added to those of document d produce a document containing no words.

Suppose $\underline{d}'$ was such a document, given by,

$$\underline{d}' = (g_1, g_2, \ldots . g_n).$$

Then adding $\underline{d}$ to $\underline{d}'$ would give,

$$\underline{d} + \underline{d}' = (f_1 + g_1, f_2 + g_2, \ldots . f_n + g_n)$$

$$= \underline{\phi}$$

$$= (0,0, \ldots . 0)$$

That is, $f_i + g_i = 0$, for $i = 1, \ldots n$.

Mathematically, then, the problem reduces to finding an inverse g_i for each natural number f_i, but this is not possible within the set of natural numbers, which itself is not a group for this very reason. To solve this we must extend our range of frequencies f_i to include negative integers. However, to speak of a word having a negative frequency of occurrence in a document is meaningless, and we cannot find the necessary documents to extend D so that its associated vector set $\underline{D}$ becomes a group, with inverses.

Hence we see that $\underline{D}$ itself does not constitute a vector space. However, it is possible to find a vector space $\underline{V}$ of which $\underline{D}$ is a subset. In fact the set of n-tuples $(a_1, a_2, \ldots . a_n)$, where a_i is any integer, positive, negative or zero, is such a vector space, and $\underline{D}$ is the restriction to members whose a_i values are positive or zero only. We shall continue to use the terminology of vector spaces for discussing document collections, but will be concerned only with the members of the subset $\underline{D}$.

II.4 Word pairs

In section II.3.1 we mentioned the loss of order of words in a vector representation. This order is important as it serves to resolve ambiguities of meaning in individual words. One possible way of incorporating word order into the vectors is by using phrases as the coordinates of the vector space. The elements of each vector in this case are the frequencies of occurrence of the phrases in the document it represents. Here we use the word 'phrase' to mean any string of words, not necessarily constituting a grammatical phrase. In fact, experiments by Weiss (26) on the use of phrases as indexing units showed that only 37% of the phrases identified were syntactically correct, when a phrase was taken to be any string of words of a particular chosen length.

II.4.1 Choice of phrases

To be complete we should include as coordinates all phrases of length 2 up to length k, where k is the length of the longest phrase found in any document. This would be prohibitively expensive in both computer space and time, and we have therefore restricted the use of phrases to those of length 2, that is, word pairs.

Every distinct pair of adjacent words found in the document collection is used as a coordinate in the vector space. The pairs are particularly useful for distinguishing the several meanings of homonyms. For example, the word 'general' as used in the phrase 'in general' can be distinguished from the same word used in the pair 'Secretary General'. In the single word scheme all occurrences of the word 'general' are assigned to the same coordinate, regardless of meaning.

Of course there will be some false combinations arising from the non-grammatical nature of the pairs. Consider, for example, the following sentence:

'In this document analysis of social security schemes takes place.'

This does not deal with the analysis of documents, despite the occurrence of the pair 'document analysis', which would be a coordinate in the word pair scheme. However, classification takes into account all pairs in the sentence, thus placing these two words into their proper context, and so minimising the effect of the false combination.

False combinations present more of a problem for searching, where only a few word pairs in a document are used for matching with any one question. To overcome these errors, experiments have been undertaken to identify the grammatical phrases, for example those by Weiss (27) and Salton (28), but the improvement in retrieval performance has not been sufficient to justify the huge increase in costs.

II.4.2 Common words in word pairs

We have the option of excluding the common words before identifying the word pairs occurring in the documents. This would seem a reasonable thing to do as we do not find the common words useful as single index terms. However, Neufeld et al (29), who use word pairs to index titles for a current awareness publication, favour the inclusion of the frequently occurring general terms which are often regarded as common words, since these can form meaningful pairs with non-common words or other common words. For example, the word 'analysis' often occurs as a general term, e.g. in the phrase 'analysis of results', but is nevertheless very important in the phrase 'cluster analysis'.

In this study we investigate both possibilities of including and excluding common words, and the results are presented in later chapters.

II.5 Conclusion

A method for representing the full text of documents as mathematical vectors has been described. Many of the mathematical properties of vectors hold true for document vectors, although a document collection by itself does not form a vector space, but is a subset of a larger set which does. The elements of the document vectors can be the frequencies of occurrence of either single words or word pairs.

The program which constructs document vectors is described in Chapter IV, and the results of using the various strategies based on single words and word pairs, with and without common words, are discussed in Chapter V. Further experiments with alternative characteristics of text as vector elements are discussed in Chapters VII and VIII.

CHAPTER III

CLUSTER ANALYSIS

The document vectors described in the previous chapter are constructed in order to facilitate the numerical manipulation of document descriptions to determine their similarity. This chapter describes the numerical methods used for classifying documents on the basis of this similarity.

III.1 Introduction

Cluster analysis is a numerical method for analysing multivariate data. It is used to group objects, each measured on a given set of variables, into classes, or clusters. Each cluster contains objects which are more similar to one another than to those objects not in that cluster.

Various other names have been given to this process, including numerical taxonomy, clumping, classification, typology, Q-analysis and unsupervised pattern recognition, which reflect the diversity of uses for the technique -

> ". . . to classify soils and diseases, politicians and plant communities, archaeological artefacts and oilbearing strata, socioeconomic neighbourhoods and psychological types, languages and television programs - to name just some of the applications."
>
> Sokal (6).

Hill (30) has even used cluster analysis to classify cocktails on the basis of their ingredients.

The first modern uses of numerical methods in classification were in the fields of biology and zoology, notably the work by Sneath (31), and Michener and Sokal (32), in 1957. A chronological chart showing the development of cluster analysis from this time is presented by Sneath and Sokal (33). Extensive reviews of past work in the subject are also given by Ball (34) and Cormack (35). These references provide comprehensive bibliographies of the development and application of cluster analysis, so further discussion is not included here except to mention the work of Jardine and Sibson (36) which is noteworthy for its rigorous mathematical approach to classification.

III.2 Methods of clustering

An excellent introductory review of specific clustering methods is given by Everitt (37), and more advanced discussions of techniques are presented by Sneath and Sokal (33). We discuss here the major differences in the various methods and refer to the more widely used ones.

We consider clustering from five different aspects. These are, whether the method is

(i) monothetic or polythetic
(ii) agglomerative or divisive
(iii) hierarchical or non-hierarchical
(iv) overlapping or non-overlapping
(v) direct or iterative.

III.2.1 Monothetic and polythetic clustering

A monothetic method produces clusters which are each defined by a unique set of attributes. The possession by an object of a given set of attributes is a necessary and sufficient condition for membership in the cluster defined by those attributes. In some cases the clusters are defined by a single attribute. The success of a monothetic method depends heavily on the choice of attributes which define clusters. Serious misclassification can arise from the wrong choice.

Monothetic methods are usually used for binary data, since decisions are based only on the presence or absence of attributes, and quantitative information is not required. It is particularly easy to construct hierarchies from monothetic clustering schemes.

In contrast to the monothetic approach, a polythetic method clusters two objects together if they have a large but unspecified number of attributes in common. Within one cluster the set of common attributes may differ for each pair of members. No one attribute is necessary or sufficient for membership in any particular cluster.

Beckner (38) has formally expressed the idea of 'polythetic classes'. He begins:

> "A class is ordinarily defined by reference to a set of properties which are both necessary and sufficient (by stipulation) for membership in the class."

The clusters produced by a monothetic method are classes in this sense.

Beckner continues:

> "It is possible, however, to define a group K in terms of a set G of properties
>
> $f_1, f_2, \ldots f_n$
>
> in a different manner. Suppose we have an aggregation of individuals (we shall not as yet call them a class) such that:
>
> 1) Each one possesses a large (but unspecified) number of properties in G
> 2) Each f in G is possessed by large numbers of these individuals, and
> 3) No f in G is possessed by every individual in the aggregate
>
> By the terms of 3), no f is necessary for membership in this aggregate; and nothing has been said to either warrant or rule out the possibility that some f in G is sufficient for membership in the aggregate."

Clusters formed by a polythetic method satisfy conditions 1 and 2, and may or may not satisfy 3. If condition 3 is not satisfied, and some attribute is possessed by all members of a cluster, this is accidental and not a condition for the choice of the cluster.

A polythetic cluster is a similar concept to the 'fuzzy set' formulated by Zadeh (39), which defines a set by a 'characteristic function'. For each member of the fuzzy set the function determines its grade of membership in the set. A characteristic function could be constructed for a polythetic cluster by defining it to be some measure of the

number of the attributes $f_1, f_2, \ldots f_n$ that each member possessed. If an object possessed none of the attributes defining a given cluster, then its value of the function would be zero and the object excluded from the cluster, in analogy with zero grade of membership in a fuzzy set, which implies 'non-membership'.

Polythetic clusters and fuzzy sets provide a means of grouping ill-defined objects. Often we cannot say definitely that something does or does not belong to a class, but merely that it satisfies some of the criteria which define the class. For objects that are ill-defined, polythetic clustering methods are more appropriate than monothetic ones, and are more likely to produce clusters that are 'natural'.

III.2.2 Agglomerative and divisive clustering

An agglomerative method begins by placing each object in its own separate cluster. At each subsequent stage, clusters are formed by taking unions of the clusters present at the previous stage. The criteria for the union of clusters are based on the possession of certain attributes by the members of the candidate clusters. The clustering is complete when all objects belong to a single cluster.

Divisive clustering operates in the opposite fashion. At first all objects belong to the same cluster. This is then divided, according to some criteria, into two or more clusters, and these are subsequently further divided until

all clusters consist of one object each. No unions of clusters are allowed at any stage.

In short, agglomerative methods successively join clusters, and divisive methods successively partition them. A hierarchy can be constructed to show the development of the classification. However, if one particular grouping of objects is required then a suitable stopping point must be chosen before reaching the state where all objects belong to one cluster, in the agglomerative case, or where they all exist as individual clusters.

Agglomerative methods are more desirable than divisive ones because they are more flexible; any two subgroups can be considered for possible amalgamation. For a set of n objects this involves at most $n(n-1)/2$ possibilities, this worst situation occurring at the first step when any two of the n objects can be combined to form a cluster. The first step in a divisive method is to divide the collection into two or more groups. For the simple case of dividing in two, this can be done in $2^{n-1}-1$ different ways, and it is impractical to consider all these possibilities, except when n is quite small. We must therefore restrict ourselves to a smaller number of choices, thus predetermining, to a certain extent, the final outcome of the divisive clustering.

Divisive methods are normally used in conjunction with a monothetic strategy, and are perhaps preferable in these cases as it is very efficient to subdivide groups on

the basis of a single attribute. The best known of the divisive methods is that of 'association analysis' used by Williams and Lambert (40).

Agglomerative methods are preferred for polythetic strategies, as they involve much less computing than divisive methods in this case. An example of an agglomerative method is the single-link cluster method to be described in section III.4.

III.2.3 Hierarchical and non-hierarchical clustering

Different authors use the word 'hierarchical' to describe different cluster schemes. Jardine and Sibson (41) mean by a hierarchical system a 'nested sequence of partitions'. This implies that the clusters at any one level in the hierarchy are disjoint. Sneath and Sokal (33) allow hierarchical schemes to include overlapping clusters, and require only that the clusters be ranked one above another.

These 'overlapping hierarchic schemes' are defined by Jardine and Sibson (41) in terms of nested partitions as follows:

A k-partition is defined to be a grouping into subsets which allows a maximum of k-1 objects in the overlap between any two subsets. An overlapping system is then considered to be a 'nested sequence of k-partitions'.

A hierarchical scheme is taken to be the case when k is equal to 1, and there is no overlap.

We will regard any classification scheme which ranks clusters as hierarchical, since the discussion applies to both overlapping and non-overlapping cases.

Hierarchical methods are useful when we wish to see how clusters are related to one another, and want to summarise the relationships between the individual objects. Non-hierarchical classifications exhibit more clearly the precise relationships between individuals, and are likely to consist of more homogeneous clusters, as the non-hierarchical methods aim to optimise the relatedness of objects in each cluster. Imposing a hierarchy weakens the homogeneity of clusters, in order to demonstrate inter-cluster relatedness.

One advantage of the non-hierarchical approach is that any object can be re-allocated at a later stage, if it is found to have been allocated to the wrong cluster initially. The assignment of an object to a cluster in a hierarchical scheme is irrevocable, as the object cannot be moved without altering the whole hierarchy. However, hierarchical methods generally require much less computing time, and so are more feasible for use with large data bases.

III.2.4 Overlapping and non-overlapping clustering

Overlapping clusters were introduced in the previous section, viz. k-partitions. Another example of clustering which produces overlap is that of 'clumping' used by Needham and Sparck Jones (42). Overlap between clusters is desirable because an object can often be associated with more than one cluster on the basis of various subsets of the attributes the object possesses. However, much more computing is involved than for non-overlapping schemes, and the results can be quite cumbersome to represent diagrammatically for any reasonable amount of overlap. For these reasons overlapping techniques are not widely used.

III.2.5 Direct and iterative clustering

The name 'direct' implies that the method proceeds directly to the end result by performing the classification process once. An iterative method repeats the process until some optimum result is achieved. An ideal iterative method would examine every possible solution and select the one which best satisfied the criteria for optimality. This is generally too large a task, and most iterative methods employ what Sneath and Sokal (33) call a 'hill-climbing technique', which attempts always to improve on the preceding result. Unfortunately this may result at some stage in a local optimum, so that it becomes impossible to proceed to

the absolute optimum without first reverting to some worse classification, and hence the local optimum must be accepted as the final result. To quote Sneath and Sokal (33):

> "The problem is like that of a man moving in a random direction trying to climb to the top of the Rocky Mountains through a dense fog: if his strategy is always to climb upward he will most likely be trapped on the peak of one of the lower ranges."

III.3 Choice of method

It is generally agreed that there is no one 'best' method of clustering, and it is not easy to choose from the multitude of methods that exist. However, certain characteristics will be desired in the resulting classification scheme, and these will determine the choice to a certain extent. For example, we may wish to impose constraints on the homogeneity of the clusters, or on the size and number of clusters produced. Consider, for example, the problem of allocating work to the various departments of an administration. There would be restrictions on the number of classes of jobs, determined by the number of departments, and possibly restrictions on the size of each class, depending on the maximum amount of work with which each department could cope. Moreover, in this example non-overlapping clusters would be required, as we would not want the same job to be done by more than one department.

Another major consideration when choosing a method is the ease with which the classification can be performed. For large data bases, some of the more complex and sophisticated methods become almost impossible computationally.

In addition to these rather vague guidelines, there have been a few attempts at formulating more rigorous criteria for a 'good' clustering method. The two major, and partly conflicting, contributions in this area have been made by Jardine and Sibson (41) and Williams et al (43). Further controversial discussions of the criteria devised by these authors are contained in Williams et al (44), Sibson (45), and Jardine and Sibson (46).

Jardine and Sibson's conditions point to 'single-link clustering' as being the most mathematically sound of the hierarchical methods. Because of these mathematical advantages and its ease of implementation, the single-link method was chosen for this study of the classification of legal documents. In the following section we describe the method in detail, in terms of the algorithm used to implement it. We also discuss the advantages and disadvantages of single-link clustering and give further reasons for its choice in this case.

III.4 Single-link clustering

The single-link clustering method (also known as nearest neighbour) produces a hierarchical, non-overlapping classification. The degree of association of an object with the cluster it joins is equal to the highest degree of association it has with any one of the existing members of the cluster. Similarly, two clusters are joined at a level of association equal to the highest degree of association existing between any pair of objects taken one from each of the two clusters.

The method operates on the values of the function which measures this degree of association between all possible pairs of objects in the collection to be clustered. We refer to this measure of association as the 'similarity coefficient', though in the discussion in section III.5 we will see that some functions actually measure the dissimilarity between objects.

In the following discussion the collection of objects to be clustered is considered to be the set of document vectors $\underline{D}$ discussed in Chapter II. The similarity coefficient is defined by a function ρ on $\underline{D} \times \underline{D}$, the set of pairs $(\underline{v},\underline{w})$ where $\underline{v}$ and $\underline{w}$ are vectors in $\underline{D}$; that is, the value $\rho(\underline{v},\underline{w})$ is the coefficient of similarity between $\underline{v}$ and $\underline{w}$.

III.4.1 Clustering algorithm

We discuss single-link clustering in terms of a direct, agglomerative algorithm, based on one by van Rijsbergen (47).

The similarity coefficients, calculated for all pairs of vectors, are first sorted into descending numerical order. These are then processed sequentially according to the following rules:

at each level $\rho(\underline{v},\underline{w})$ of the similarity coefficient

(1) if both $\underline{v}$ and $\underline{w}$ belong to the same cluster, then take no action

(2) if neither $\underline{v}$ nor $\underline{w}$ belongs to a cluster, then form a new cluster at level $\rho(\underline{v},\underline{w})$ consisting of $\underline{v}$ and $\underline{w}$

(3) if one of $\underline{v}$ or $\underline{w}$ belongs to a cluster but the other does not, then join the other to this cluster at level $\rho(\underline{v},\underline{w})$

(4) if both $\underline{v}$ and $\underline{w}$ belong to distinct clusters, then join these two clusters at level $\rho(\underline{v},\underline{w})$.

Initially we may regard each vector $\underline{v}$ as a cluster containing itself alone, and eventually we reach some level ρ at which the whole of $\underline{D}$ forms one cluster. The hierarchy of clusters produced can be represented by a tree diagram called a dendrogram, an example of which is given in the following section.

III.4.2 Worked example

In the following example we show how single-link clustering operates on a given matrix of similarity coefficients, and draw the resulting dendrogram.

Consider that we have five objects to be clustered, labelled 1 to 5. Let the coefficients of similarity between pairs of objects be given by the matrix

	1	2	3	4	5
S = 1	1.00	0.45	0.90	0.30	0.60
2	0.45	1.00	0.10	0.75	0.40
3	0.90	0.10	1.00	0.50	0.55
4	0.30	0.75	0.50	1.00	0.20
5	0.60	0.40	0.55	0.20	1.00

Note that the coefficient of similarity between an object and itself is 1.00, the maximum value of the coefficient in this case, indicating total similarity. Note also that the similarity function is 'symmetric', that is $\rho(\underline{v},\underline{w}) = \rho(\underline{w},\underline{v})$. Thus we need use only the portion of the matrix enclosed in the triangle.

Sorting the coefficients into descending order gives

Coefficient	Object numbers
0.90	1,3
0.75	2,4
0.60	1,5
0.55	3,5
0.50	3,4
0.45	1,2
0.40	2,5
0.30	1,4
0.20	4,5
0.10	2,3

The first cluster formed is (1,3), at the highest level 0.90. At the next level 0.75, the pair (2,4) becomes a second cluster. Stepping on we find the next pair is (1,5), and since 1 already belongs to a cluster, 5 joins this cluster too, giving (1,3,5). The next pair 3 and 5 already belong to the same cluster so no action is required, and we proceed to level 0.50, which is the coefficient of similarity between 3 and 4. Each of these belongs to a distinct cluster so we join these clusters at level 0.50. All five objects now form a single cluster, and the clustering is complete.

The dendrogram representation of this clustering is shown in Figure III.4.

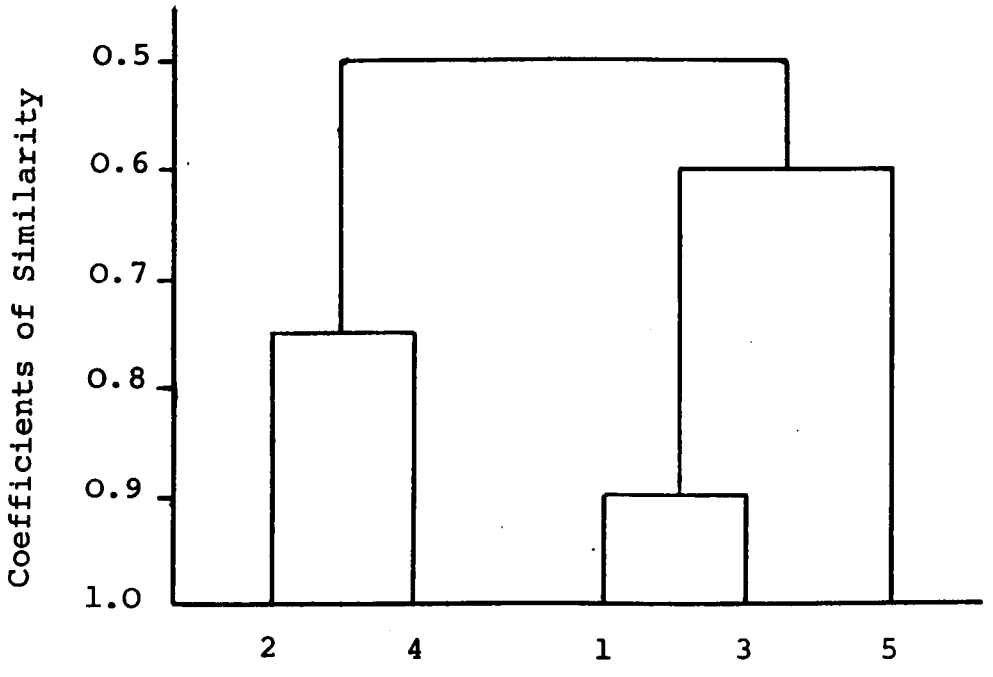


FIGURE III.4 Dendrogram showing single-link clustering

The levels in the dendrogram at which new objects join clusters and at which two clusters merge are called 'splitting levels'. Clusters formed at high splitting levels, that is, at high values of the similarity coefficient, are the most homogeneous. At lower levels the inter-cluster relationships are demonstrated.

III.4.3 Advantages of single-link clustering

The merits of the single-link method have been discussed by Jardine and van Rijsbergen (19) in relation to document retrieval systems.

These advantages are:

1. The clustering depends only on the rank-ordering of the similarity coefficients. This is important since there is often no statistical justification for the choice of the similarity function used. For example, the value of the function could be multiplied by some scalar value without altering the order of the coefficients or the resulting classification.
2. The clustering is stable in the sense that small errors in the values of the coefficient effect only small changes in the clustering. Errors in the classification resulting from miscalculation or wrongly-assigned attributes will be very minor.
3. The clustering is unlikely to change drastically when new objects are introduced into the system.

III.4.4 Disadvantages of single-link clustering

Despite its mathematical advantages and computational ease, the single-link method is not perfect in every way. It is mostly criticised for its 'chaining' effect. This is the term applied to the long straggly clusters consisting of objects linked by chains of intermediates. The dendrogram in this case is not very informative, because it does not display how the intermediate chains are built up; that is to say, each subsequent object which is 'chained' to a cluster

is not necessarily linked to the previously chained object, but may be linked to an object which joined the cluster much earlier.

Chaining is a disadvantage if homogeneous compact clusters are required, but these may not necessarily exist so that in some cases a 'straggly' chained classification may be a true representation of the structure of the data. As Jardine and Sibson (41) point out, it is rather misleading to call chaining a defect of single-link clustering; it is simply a description of what the method does.

Most hierarchical clustering methods produce relatively non-homogeneous clusters, and efforts have been made to overcome this problem. Wishart (48), for example, has attempted to reduce chaining in single-link clustering. However, it is advisable to use a non-hierarchical method if compact clusters are essential.

III.4.5 Choice of the single-link method for classifying legal documents

The choice of a clustering method depends mainly on the uses to which the classification is to be put. The possibilities for using a classification scheme in a document retrieval system were discussed in Chapter I, and it is with these uses in mind that we consider the choice of a method for classifying legal documents.

The single-link method was chosen mainly for its hierarchical nature. A hierarchy of documents is especially desirable for improving the recall of a search; as each cluster of documents is exhausted there is a natural progression to the next related cluster in the hierarchy, extending the search. The splitting levels in the hierarchy can be used to determine cut-off points for searches.

It is no disadvantage that hierarchical methods do not produce highly homogeneous clusters, as these would not provide the improved recall we desire. The members of a compact cluster would probably all be retrieved initially. We are concerned with finding those documents which contain the topic of interest as a peripheral, and which would not normally be recalled by a specific request for that topic. Moreover, very efficient strategies can be devised for searching hierarchies.

Using a classification scheme for efficient storage of data for retrieval was another of the possibilities discussed in Chapter I. In this case a non-overlapping scheme is desirable in order to avoid having to store documents more than once. The linear arrangement of documents on the horizontal axis of the dendrogram suggests a means of implementing the storage scheme.

The single-link method is both hierarchical and non-overlapping. Moreover it is very easy to implement even for large data bases, and was therefore chosen to classify a large collection of legal documents.

A description of the particular similarity coefficient used for the classification is included in the following discussion of measures of association.

III.5 Measures of association

In the previous discussions on clustering we referred to a measure of association between objects to be clustered, which we called a similarity coefficient. There are many different functions which can be used to calculate these values, and we discuss these in general and give some of the better known examples.

Sokal and Sneath (49) have reviewed association measures extensively, and this review has been added to in Sneath and Sokal (33). They divide the various measures into four groups. These are:

1. Distance coefficients.
2. Association coefficients.
3. Probabilistic coefficients.
4. Correlation coefficients.

Examples of these measures are given in the following sections, in terms of document vectors where appropriate.

III.5.1 Distance coefficients

Distance coefficients measure the distance between the two points in metric space which represent the objects being compared.

An example is the Euclidean distance

$$D(\underline{v},\underline{w}) = \left(\sum_{i=1}^{n} (f_i - g_i)^2 \right)^{\frac{1}{2}}$$

where $\underline{v} = (f_1, f_2, \ldots . f_n)$

$\underline{w} = (g_1, g_2, \ldots . g_n)$

are vectors in an n-dimensional vector space.

The disadvantage of the Euclidean distance is that it is greatly affected by the scale factor for each dimension of the attribute space, especially when the dimensions are not commensurate in units of measurement; for example, one dimension measured in feet, another in seconds.

Distance coefficients are all metrics. That is, they all satisfy the following conditions:

(a) $D(\underline{v},\underline{w}) \geq 0$, and $D(\underline{v},\underline{w}) = 0$ if and only if $\underline{v} = \underline{w}$.

(b) $D(\underline{v},\underline{w}) = D(\underline{w},\underline{v})$.

(c) $D(\underline{v},\underline{w}) + D(\underline{w},\underline{z}) \geq D(\underline{v},\underline{z})$.

The more dissimilar two objects are, the greater the distance is between them. Hence distance coefficients are dissimilarity coefficients. Clustering algorithms operate equally well on both similarity and dissimilarity coefficients. The description of single-link clustering given in section III.4 applies in reverse to dissimilarity coefficients. For example, the algorithm would process the coefficients in ascending order, and the more homogeneous clusters would be those formed at a low splitting level.

III.5.2 Association coefficients

Association coefficients are used for measuring the similarity between objects described by two-state or multi-state attributes, in terms of presence or absence of these attributes.

In the following example I and J are objects in a system described by a set of n attributes.

Let,

a = the number of attributes present in both I and J

b = the number of attributes present in I and not in J

c = the number of attributes present in J and not in I

d = the number of attributes absent from both I and J.

(Note that a + b + c + d = n.)

Two of the simpler association coefficients are:

$$\text{Sneath's} \quad S(I,J) = \frac{a}{a + b + c}$$

$$\text{Dice's} \quad S(I,J) = \frac{2a}{2a + b + c}$$

Neither of these two coefficients takes into account those attributes which are absent from both I and J. An example which does include the factor 'd' is the Simple Matching Coefficient

$$S(I,J) = \frac{a + d}{a + b + c + d}$$

III.5.3 Probabilistic coefficients

Probabilistic measures take into account the distribution of the attributes over all the objects in the collection, on the assumption that agreement of a pair of objects on a rare attribute contributes more to their similarity than agreement on a frequently occurring attribute. An example is Goodall's probabilistic similarity index (50). However, this and other such measures are extremely complex and involve large amounts of computation, and are therefore not suitable for sizeable data bases.

III.5.4 Correlation coefficients

The best known of the correlation coefficients is the Pearson product moment correlation coefficient. This is given by the expression:

$$S(\underline{v},\underline{w}) = \frac{\sum_{i=1}^{n} ((f_i - \bar{f})\ (g_i - \bar{g}))}{\left(\sum_{i=1}^{n} (f_i - \bar{f})^2 \sum_{i=1}^{n} (g_i - \bar{g})^2 \right)^{\frac{1}{2}}}$$

where $\underline{v} = (f_1, f_2, \ldots . f_n)$

$\underline{w} = (g_1, g_2, \ldots . g_n)$

and $\bar{f}$ and $\bar{g}$ are the mean values of the attribute values $f_1, f_2, \ldots . f_n$, and $g_1, g_2, \ldots . g_n$, respectively.

By standardising the values of the attributes such that the mean values $\bar{f}$ and $\bar{g}$ are zero, the coefficient reduces to:

$$S(\underline{v},\underline{w}) = \frac{\sum_{i=1}^{n} (f_i g_i)}{\left(\left(\sum_{i=1}^{n} f_i^2 \right) \left(\sum_{i=1}^{n} g_i^2 \right) \right)^{\frac{1}{2}}}$$

This is the well known cosine coefficient which is also widely used with unstandardised data. It is useful for its geometrical interpretation; $S(\underline{v},\underline{w})$ is the cosine of the angle between the two vectors $\underline{v}$ and $\underline{w}$ in n-dimensional space.

That is, in Figure III.5, $S(\underline{v},\underline{w}) = \cos\theta$. The closer the two vectors are, the smaller the angle is between them, giving a higher value for the similarity coefficient. The numerator of the cosine coefficient, $\Sigma f_i g_i$, is the scalar product of the vectors $\underline{v}$ and $\underline{w}$. The factors $(\Sigma f_i^{\,2})^{\frac{1}{2}}$ and $(\Sigma g_i^{\,2})^{\frac{1}{2}}$ in the denominator are the respective lengths of the two vectors.

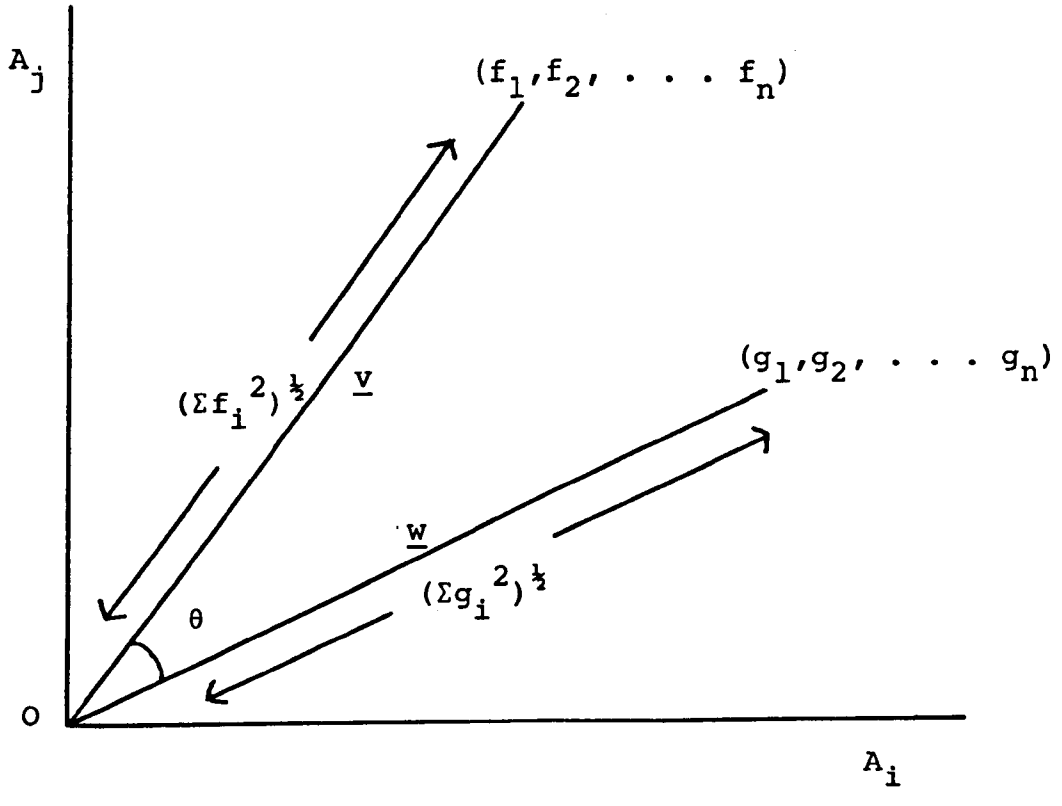


FIGURE III.5 Geometrical interpretation of the cosine coefficient of similarity between two vectors

III.6 The cosine coefficient

The coefficient used here for clustering legal documents is the square of the cosine coefficient. We denote this particular coefficient by ρ, which is given by:

$$\rho(\underline{v},\underline{w}) = \frac{(\sum_{i=1}^{n} f_i g_i)^2}{(\sum_{i=1}^{n} f_i^2)\ (\sum_{i=1}^{n} g_i^2)}$$

The f_i and g_i are the frequencies of the words occurring in the documents represented by the two vectors $\underline{v}$ and $\underline{w}$. In practice the factor $\Sigma f_i g_i$ is the sum of the products of the frequencies of just those words which occur in both $\underline{v}$ and $\underline{w}$. The factors Σf_i^2 and Σg_i^2 are the sums of the squares of the frequencies of the words occurring in $\underline{v}$ and $\underline{w}$ respectively.

The maximum value of ρ is 1, when both documents contain exactly the same words. When the two documents have no words in common we get the minimum value of zero. We never get negative values for this coefficient, because the angle θ can not be greater than 90°. This is because negative frequencies of occurrence never arise, and all the vectors lie in the positive sector of the n-dimensional space.

The cosine coefficient was chosen for its geometrical interpretation, and the square of the cosine used to simplify calculations. Squaring the coefficient makes no

difference to the single-link clustering obtained, since the result depends only on the numerical order of the coefficients, and the square of the cosine is jointly monotonic with the cosine;

$$\text{i.e. if } \cos\theta_1 \geq \cos\theta_2$$

$$\text{then } \cos^2\theta_1 \geq \cos^2\theta_2$$

for our range of values for θ lying between 0° and 90°.

The cosine coefficient has been used extensively for document retrieval by Salton (28) in the SMART system.

III.6.1 Some properties of the cosine coefficient

Because of the factors in the denominator, the cosine coefficient tends to give low values for long documents, except where they match on a large number of terms. However, it is important to include the frequencies of all unmatched terms in the calculation, since, for example, one word out of a possible hundred co-occurring in a pair of documents is less significant than one word out of ten co-occurring. In this way we can distinguish a document containing a particular topic as its main theme from one which contains the topic as a peripheral idea only, in addition to many other topics.

Another feature of the coefficient is that multiplying all the elements of one vector by some constant has no effect on the values of similarity between it and other vectors, as this constant factor cancels top and

bottom from the coefficient. This means, for example, that two documents may appear totally similar, having a coefficient of 1, when in fact one document contains the same words as the other, but twice as many times. However, from a semantic viewpoint, it is unlikely that the longer document contains any more ideas since it uses no different words, and the two documents can be taken to be as similar as two documents containing the same words with exactly the same frequencies.

III.7 Evaluating a classification scheme

The previous sections have described the tools and techniques for generating a classification scheme by cluster analysis. Having constructed a classification we are faced with the problem of evaluating it. The difficulties are perhaps best summed up by Rubin (51):

> " We do not have a clearly defined mathematical problem unless we can evaluate the degree of organization we have achieved by the resulting set of groups. If we had a function to measure the value of a given grouping (or, at least, a way of deciding which of two groupings is better), we could in theory examine every possible way of grouping the objects, and select the best (the one which optimizes the function.) Without such a criterion function we are in a much more ill-defined situation in which we can use procedure after procedure, with no objective method of deciding which of these different results is best.

> The problem of finding the 'best' grouping of a set of data is not well-defined unless one can specify what makes one 'better' than another. To specify this requires an analysis of the purpose of performing the categorization."

To evaluate a classification scheme, then, we examine how well it performs the function for which it was constructed. If this is satisfactory we can accept the classification, otherwise we must look for another which performs the task 'better'.

We cannot judge the success of a classification on the mere size, shape and number of clusters obtained. There may be few clusters present in the data, so a classification which formed only a few would be good in this case. There may of course be some internal criteria, such as size, which we wish the clusters to satisfy. These should be used as guidelines for choosing the method of clustering originally, and evaluating the clusters on the basis of these requirements would be more a test of how well the clustering method had been chosen. That is, if we make the best choice then we will necessarily get the best classification from the point of view of these criteria.

It is desirable to have a numerical method for evaluating classification, and one has been developed by Sokal and Rohlf (52). This involves calculating what is called the 'cophenetic correlation coefficient', which compares the original similarity matrix with the matrix

of coefficients implied by the clustering. For example, for a clustering method yielding a dendrogram, the new coefficient of similarity between a pair of objects, implied by the classification, would be the splitting level at which the two objects first belong to the same cluster. This cophenetic correlation coefficient is a measure of how well the clustering represents the original matrix of similarity coefficients. Its value can be used to compare two different clusterings of the same data, or some acceptable average value can be established by which to judge a single classification.

III.8 Conclusion

In this chapter we have discussed the various characteristics of cluster analysis, and have described the single-link cluster method which has been chosen for this study because of its mathematical advantages, hierarchical output and ease of implementation. The cosine coefficient used here to measure the similarity between legal documents has been described with reference to its geometrical interpretation, as part of a general discussion of association measures.

The programs used to calculate cosine similarity coefficients and to perform single-link clustering on legal documents are described in the following chapter, Chapter IV. Some experimental classifications are presented and discussed in Chapters V, VII and VIII.

CHAPTER IV

COMPUTER PROGRAMS

IV.1 Introduction

This chapter describes a suite of five computer programs which are used to generate a single-link clustering of a collection of documents. The clustering may be based on the occurrence of single words in the documents, or alternatively on the occurrence of pairs of words. A sixth auxiliary program produces a concordance of words in the document collection classified.

The programs are all written in FORTRAN for ease of portability, although some of the character handling routines are machine-dependent. The suite has been implemented on two installations, the ICL 4130 at the University of Kent, and the IBM 370/165 at UKAEA, Harwell.

The IBM machine was used for classifying a large collection of documents. This installation allows large amounts of core to be used cheaply, but charges heavily for computing time. Hence the programs were designed to make maximum use of the core available, in order to keep down the processing time. On the smaller ICL machine the same programs were able to process only a small amount of data. Larger document collections could be classified by small machines if more use were made of backing store, and

suggestions are given for modifications which might be made to the programs for this purpose.

IV.2 The dictionary

The first program in the suite constructs a dictionary of all the distinct words occurring in the collection of documents to be classified, and converts the text to what is known as 'word number' form, whereby each distinct word is represented by a unique integer number.

IV.2.1 Input and output

The main input to the program is the full text of the documents. There may also be some preliminary input, consisting of words which are to be regarded as 'common words', prepositions and conjunctions for example. The words are entered into the dictionary before the text of the documents is processed.

On a single pass of the data, the program assigns to each distinct word a unique integer, which will represent that word in the subsequent programs. These numbers are assigned sequentially so that the first word processed becomes number 1, the second distinct word, number 2, and so on. Subsequent occurrences of a word to which a number has already been assigned is given that same number.

'Common words' are represented by negative integers, and these are assigned sequentially beginning with -1. We refer to the word assigned number i as 'word i'. In the terminology of Chapter II word i is the attribute A_i.

As each word is identified , the dictionary in its present state is consulted to see if the word has occurred before. If it has then the count of its frequency of occurrence is updated, otherwise this new word is assigned the next positive integer and a new dictionary entry is created for it. At the same time, a file of the text in word number form is created. Each word of the text is replaced by its corresponding word number, which has been either newly assigned or retrieved from the dictionary.

The output from the program consists of the text in word number form and the dictionary. The dictionary consists of several arrays whose structure and function are described in section IV.2.3. These arrays are processed at the end of the program to produce a tabulated dictionary listing the words in alphabetical order with their word numbers, frequencies of occurrence and the number of documents in which they each occur. Finally the dictionary arrays are stored externally. They are not required for the subsequent programs in the clasification suite, but are used in the concordance program described in section IV.7 and in the search system discussed in Chapter IX.

IV.2.2 Definition of a word

A word is defined to be a string of alphabetical characters delimited by any of the following punctuation characters:

space, full stop, comma, colon, semi-colon, opening bracket, closing bracket, apostrophe.

By this definition words containing apostrophes, such as "Council's" and "l'Europe", become separated into two words, "Council" and "s", "l" and "Europe", and the single letter words are normally regarded as common words. Hyphenated words, e.g. "Secretary-General", are treated as single words, including the hyphen as a letter.

In addition, strings of four numeric digits delimited by any of the above punctuation characters are regarded as words. Such 'words' represent year numbers and are therefore semantically important. Strings of numbers of length other than four are ignored.

IV.2.3 Structure of the dictionary

(a) Binary tree

The dictionary is stored alphabetically as a binary tree; the left branch from any word leads to the word alphabetically preceding it, the right branch to the word following it.

The tree is stored as three one-dimensional arrays. The first, which we refer to as the 'tree array', holds the records which form the dictionary entries. These are variable length records, stored sequentially in the array as each new word is entered in the dictionary. Records in the tree array take the form shown in Figure IV.2(a). In this figure

A = word number assigned to word

B = number of characters in word

C = frequency of occurrence of word in document collection

D = reference number of document in which latest occurrence of word was found

E = number of different documents in which word occurs.

The shaded area contains the actual characters of the word. This area consists of a variable number of locations, up to a maximum specified by the program. A, B, C, D and E occupy one storage location each.

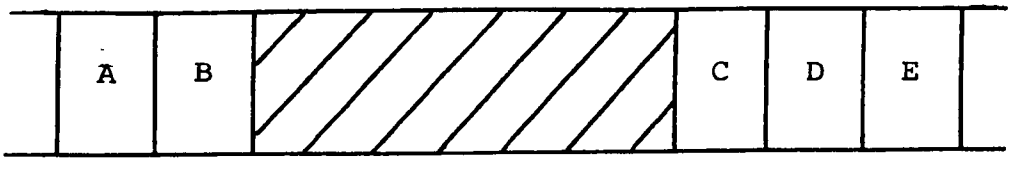


FIGURE IV.2(a) Dictionary record in tree array

The alphabetical order of the tree is maintained by the other two arrays which we call the 'left-pointer' and 'right-pointer'. Each record in the dictionary has one associated element in each of these arrays. The left-pointer element points to the record in the tree array which alphabetically precedes the given word, the right pointer element points to the record which follows.

Because the tree is constructed sequentially from the text, the word alphabetically following a given word may be found on a right branch from the given word, or, alternatively, it may constitute a node further up the tree, if it occurred in the text before the given word. In this case we must backtrack to reach the next word in order. The right pointer must be able to distinguish the two conditions, and to do this the value of the right pointer is set positive if it points to a branch, negative if it backtracks.

The left pointer of a record having no more predecessors is set to zero. Likewise the right pointer of the very last record in the alphabetical list is zero.

In the following example, Figure IV.2(b) illustrates the structure of the binary tree representation of the phrase:

"European Convention on the Transfer of Corpses"

In the diagram, backtracks are indicated by dotted lines. The word 'European' forms the root of the tree as it is the first word found in the text. It is preceded by 'Convention'

which is the first word in the alphabetical list. 'Corpses' has no right branch because the words which follow it alphabetically also follow the word 'European' and occur on its right branch. Thus 'Corpses' backtracks to 'European' which immediately follows it. The word 'Transfer' ends the alphabetical list, having neither left nor right branches and no backtrack.

This method of storing the dictionary was suggested by a sorting method called the 'Monkey-Puzzle sort' described by Day (53).

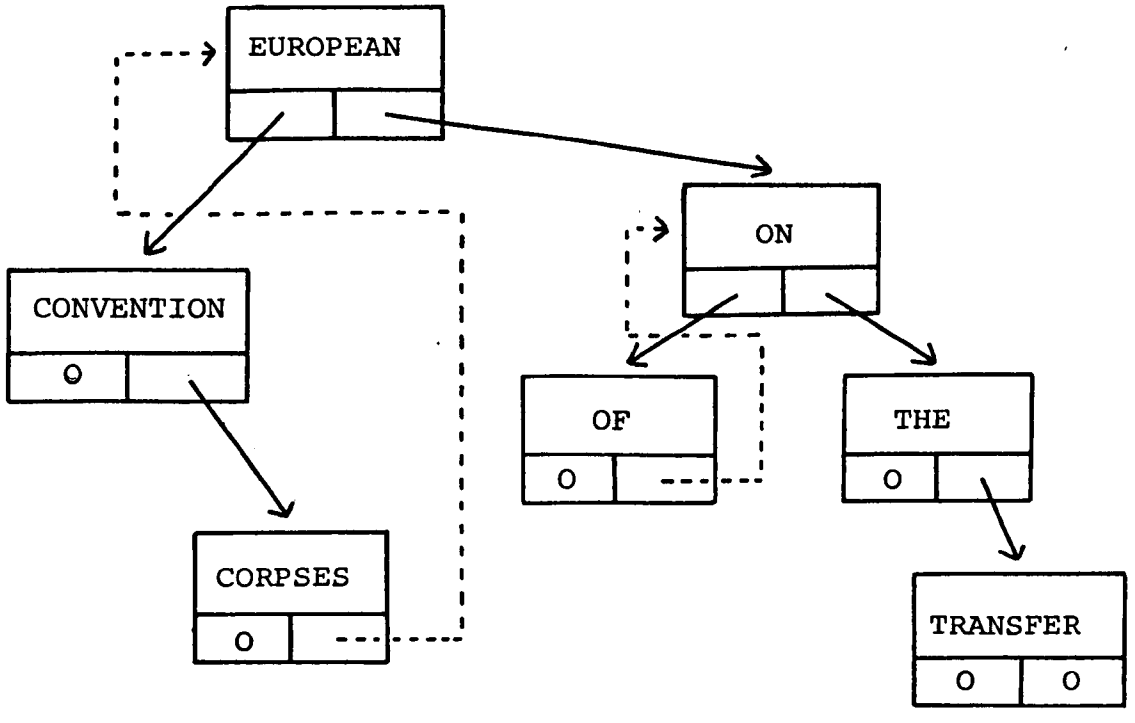


FIGURE IV.2(b) Binary tree arrangement of alphabetical word list

(b) Accessing the tree by the characters of a word

For each word identified in the input text, the dictionary must be consulted to retrieve the corresponding word number, or if the word is not already in the dictionary, to create a new record for it. We need, therefore, to access the tree array using the characters of the given word. We could search the tree alphabetically, comparing the given word with each word in the dictionary, and in fact this must be done for each new entry so that the tree pointers can be set correctly. However, we can find out directly whether or not a word is in the dictionary already, by using a 'hashing' technique which provides a pointer to the correct entry in the tree if it exists. Thus for words already in the dictionary we avoid the time-consuming alphabetical search.

Hashing, or scatter storage as it is sometimes called, is a process which transforms the internal representations of the characters of a word into some single number, the 'hash value', which is used as an index to the tree array. The function used for the transformation must always result in the same hash value for any given word. It may sometimes give the same hash value for two or more different words. This situation is allowed for, though it is desirable to choose a function which minimises the number of words hashing to the same value.

The value of the hash function is used to address an array called the 'hash table', which contains the starting addresses of the records in the tree array. Figure IV.2(c) shows the relation between the word, the hash table and the tree array. In the event of two words hashing to the same value the next free location in the hash table is used to store the address of the second word, and so on for subsequent 'crashes'.

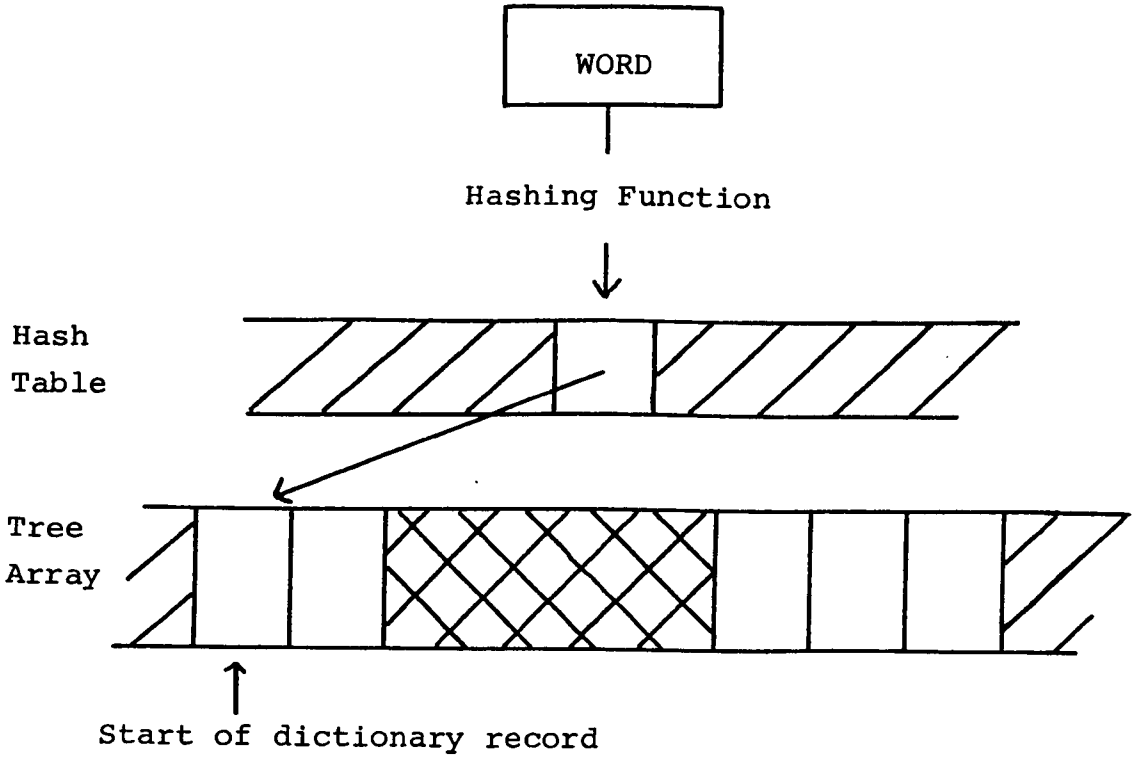


FIGURE IV.2(c) Accessing tree array by hash value

The value of the particular hashing function used in the dictionary program is calculated as follows:

The computer words occupied by the characters of the word are numbered sequentially beginning with 1 for the first computer word. The numerical value of the contents of each computer word is then divided by 2^n, where n is the number assigned to that computer word. The results of these divisions are summed to give a hash value. For example, in the case of the IBM 370 and ICL 4130 implementations each computer word holds four characters, so that the hash value is given by the value of the first four characters divided by 2 plus the second four characters divided by 4 and so on.

The division algorithm prevents overflow when adding the large numbers which represent the packed characters. In the worst possible case, if each computer word contained the maximum representable number, then the hash value would be the sum of a geometric series

$$ax + ax^2 + \ldots . ax^n$$

where a = maximum value which can be held in one computer word

$x = \frac{1}{2}$,

and the result would approach but never reach the maximum value.

The result of the above calculation will not necessarily lie within the bounds of the hash table, and is therefore normalised to do so. We take the modulo of the value with respect to the length of the hash table, and if the result is less than or equal to zero, add to it the length of the table.

There are many different ways of constructing a hashing function, and it would no doubt be possible to find one which is more efficient than the one described, in terms of computing involved and the number of collisions occurring. However, efficiency of hashing is not very important in the clustering suite, as the dictionary program which uses hashing is executed once only for any one classification of a given document collection, and the dictionary then no longer required. If, on the other hand, the dictionary were to be used continually for information retrieval, the hashing function might have to be chosen more carefully, as consultation of the dictionary forms a major part of all searches. Various hashing algorithms are described and compared by Lum et al (54), who also provide guidelines for choosing a method. Lowe (55) considers the effect different hashing functions have on retrieval performance.

The hash value is also used to access the left and right pointer arrays. That is, the pointers for a given word are contained in the locations addressed by the hash value for that word.

(c) Accessing the tree by word number

Two further arrays complete the dictionary file. These allow the tree array to be accessed by the word number assigned to a word. Two arrays are needed, one for positive word numbers and one for the negative numbers. In the 'positive' array the ith location contains the starting address of the record in the tree array for word i. Similarly, the ith location in the 'negative' array contains the corresponding information for word -i. Hence given any word number we can find from the tree the word which that number represents. Figure IV.2(d) shows the relation between these pointer arrays and the tree array.

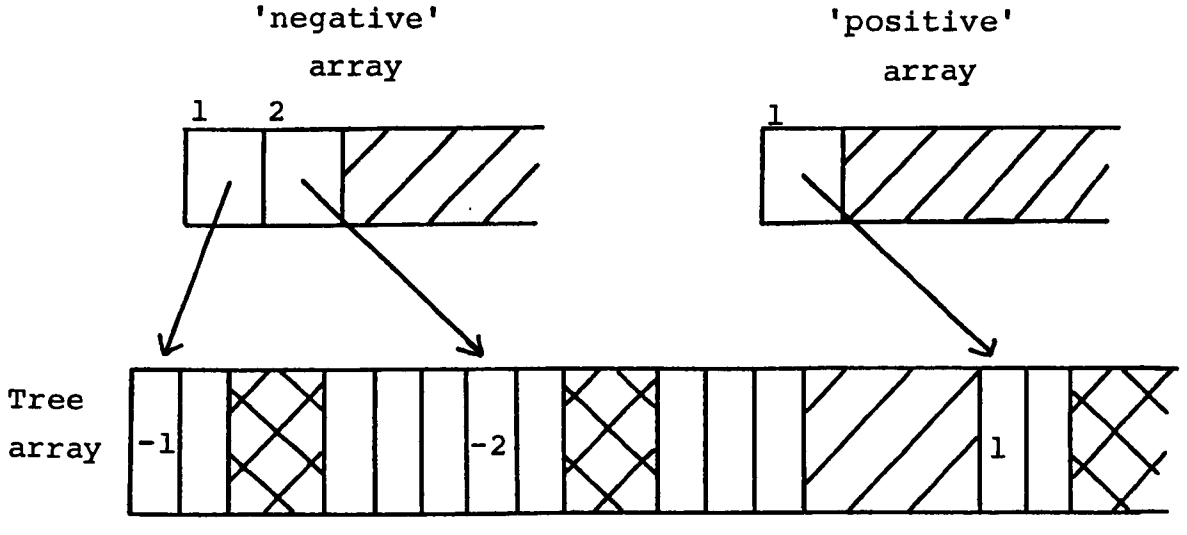


FIGURE IV.2(d) Accessing the tree array by word number

IV.2.4 The tabulated dictionary

When the dictionary is complete it is arranged into a table which lists all the distinct words in the document collection in alphabetical order, together with their word numbers, frequencies of occurrence in the document collection, and the number of different documents in which each occurs. Some common words may have been input initially which do not occur at all in the documents, but will have been entered in the dictionary. These are not listed in the table.

As each word is processed its number is entered into an array which eventually contains all the word numbers stored sequentially in alphabetical order of the words they represent. This array, together with the word number indexes described in section IV.2.3 (c), can be used to retrieve the words from the tree in alphabetical order, as an alternative to using the left and right pointers.

IV.2.5 Modifications for large data bases and small machines

All the dictionary arrays are held in core, and require a large amount of space for a sizeable data base. If a limited amount of core is available, the tree array, which is by far the largest, could be divided into small blocks and held on backing store. The appropriate block would be read in whenever the tree was to be accessed. This would, of course, slow down processing time markedly as the tree array is heavily used.

IV.3 Document vectors

This second program constructs a vector representation for each document as described in Chapter II. That is, the elements of the vectors contain the frequencies of occurrence of the words occurring in the documents. The text in word number form produced by the dictionary program is input to the vector program. The structure of the vectors output is described in the following section.

IV.3.1 Structure of document vectors

Each vector can be represented internally by a one-dimensional array whose ith location contains the frequency of occurrence of word i in the document which that vector represents. The length of the array is equal to the total number, n, of distinct words which occur in the document collection.

As each word number i is encountered on input, the frequency count in location i of the vector array for the current document is increased by one. When the vector for a given document is complete, it is reduced to a compressed form for external storage. Compression is necessary because the original array is usually sparsely populated, since only a small proportion of the n distinct words occur in any one document, and it is a waste of space to store the zero elements.

The compressed vector is stored as two one-dimensional arrays. The first contains the word numbers of just those words which actually occur in the given document. These numbers are stored in ascending numerical order. In the other array the corresponding locations contain the corresponding frequencies of occurrence. In the first two locations of the word number array we store the document reference number and the total number of non-zero elements in the vector, that is, the number of distinct words in the document. The document number is also stored in the first location of the frequency array, and the second contains the square of the geometrical length of the vector, which in this case is the sum of the squares of the frequencies.

Figure IV.3 illustrates the vector arrays for a document numbered 3, which contains nine distinct words, including word 1 three times, word 3 twice and so on.

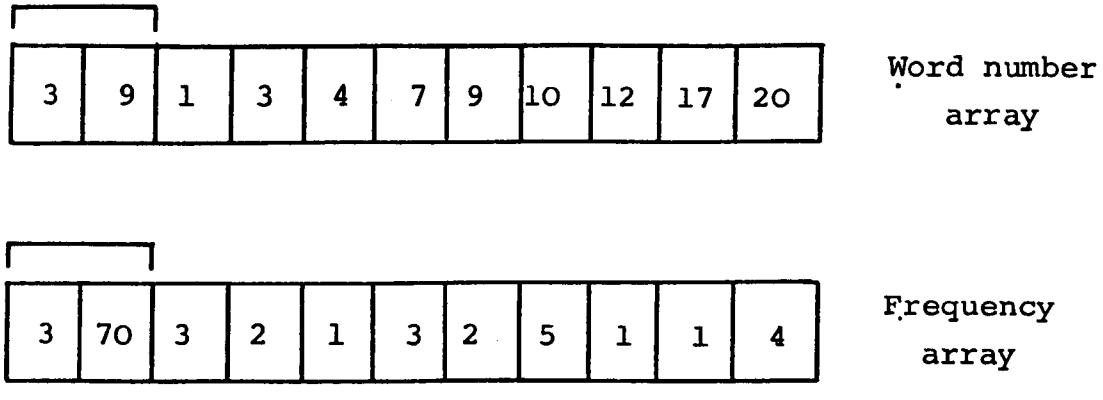


FIGURE IV.3 Compressed form of document vector

Note that in the above description of vectors we have assumed that the common words have been excluded from the representation. However, in some circumstances we may wish to include these words, and then we need a second working array in core whose length is equal to the number of distinct common words selected. In this array we store the frequencies of occurrence of word -i in location i. The contents of this array can be combined with the other working array to produce the two arrays forming the compressed vector.

IV.3.2 Modifications for large data bases and small machines

The number of distinct words, n, can be very large for data bases dealing with a wide variety of topics, and there may be insufficient room in core to hold the working vector array of length n. In this case the array could be split into blocks of suitable size and held on backing store. The appropriate block would be read in for updating frequency counts. Finally, the blocks would be read in sequentially to extract the information for the compressed vector arrays.

Alternatively, the vector could be constructed in a compressed form, with elements assigned to the word number and frequency arrays sequentially. The numerical order of the word numbers would be maintained by left and right pointers similar to those used in the dictionary program.

The pointers for the word number stored in location i of the vector array would be stored in the ith location of each of the pointer arrays. The tree system would be searched for each word number input, and if found its frequency updated. Otherwise a new entry would be made in the vector. The final compressed form of the vector would be read from the tree in numerical order using the pointers.

IV.4 Similarity coefficients

In section III.5 we discussed measures of association, and in particular the cosine similarity coefficient which is used in this study. To reiterate, the coefficient of similarity between two vectors

$$\underline{v} = (f_1, f_2, \ldots . f_n)$$

$$\underline{w} = (g_1, g_2, \ldots . g_n)$$

is given by

$$\rho(\underline{v},\underline{w}) = \frac{(\sum_{i=1}^{n} f_i g_i)^2}{(\sum_{i=1}^{n} f_i^2)\ (\sum_{i=1}^{n} g_i^2)}$$

IV.4.1 Calculation of coefficient

Each of the two factors in the denominator of the coefficient is the sum of the squares of the frequencies for one of the vectors, that is, the value stored in the 2nd

location of the appropriate frequency array. Thus these factors are already available for use in the calculation and merely have to be fetched from the array.

The term $f_i g_i$ in the numerator is non-zero only when word i occurs in both documents. The two word number arrays are scanned for these co-occurring words, and for those found the corresponding values in the frequency arrays are multiplied together and added to a cumulative sum. The final sum is squared and divided by the appropriate denominator factors to give the coefficient of similarity.

IV.4.2 Input and output

The pairs of arrays representing document vectors constitute the input, and are all held in core together, so that each is readily available for comparison with the others.

The program produces three one-dimensional arrays, each of length $m(m-1)/2$, where m is the number of documents, this length being the number of different unordered pairs which can be selected from the m documents. Two of the arrays contain document reference numbers, those stored in corresponding locations constituting a pair. The coefficient of similarity between such a pair is stored in the corresponding location of the third array. Figure IV.4 demonstrates this arrangement for four documents; the coefficient of similarity between documents 1 and 2 is 0.5 and so on.

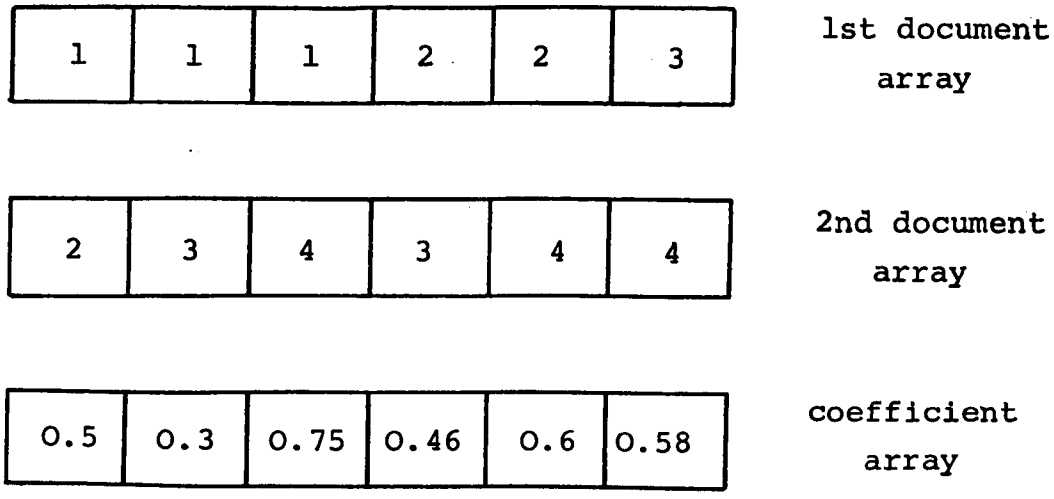


FIGURE IV.4 Arrangement of coefficients in core

These arrays are sorted into descending order of the similarity coefficients and stored externally, in the form of fixed length records, each containing three elements taken from corresponding locations in the three arrays. That is, each record consists of two document numbers together with their coefficient of similarity.

IV.4.3 Modifications for large data bases and small machines

In some cases the core storage may not be large enough to hold all the document vectors. These could be arranged in blocks stored externally and processed as follows. Read in the first block and calculate the coefficients for all pairs of documents in this block.

Keeping this block in core, read in each of the other blocks in turn, and calculate the coefficients between each of the documents in the first block and each in the new block. Now remove the first block entirely and treat the second block similarly; that is, calculate coefficients between all pairs within the block, and then compare with all remaining blocks. Repeat this process until the last block remains in core, and calculate coefficients for all pairs within this block. In this way only two blocks need be held in core at any time, and the size of the blocks can be chosen appropriately.

IV.5 Clustering

The fourth and final program in the classification suite performs the algorithm which produces a single-link clustering of the documents. Single-link clustering was discussed in detail in section III.4 in the previous chapter, and the particular algorithm used here is the one described in section III.4.1.

IV.5.1 Input and output

The similarity coefficients are processed sequentially in descending numerical order. They can be read into core at the start of the program, or if space is limited, read in one at a time as required by the program.

The output consists of a printout of the document numbers which form clusters at each level of the similarity coefficient.

IV.5.2 Clustering subroutine

The program is based on a FORTRAN subroutine written by van Rijsbergen (47). The main program controls the input of similarity coefficients, and calls this subroutine at each change in value of the coefficient.

In reference (47) the algorithm is described in terms of a dissimilarity coefficient. However, it works equally well for similarity coefficients, the main difference being the order in which the coefficients are input. The subroutine prints out clusters at each change in the value of the coefficient, and it is irrelevant whether this constitutes an increase or decrease in value.

The printout from the clustering program is used to construct a dendrogram representation of the clusters, as discussed in section III.4.2.

IV.6 Classification based on word pairs

The suite of programs described above produces a classification of documents based on the occurrence of single words. In section II.4 we discussed how we might

use pairs of words as a basis for classification. To do this we need to convert the text of the documents into 'word pair number form', analagous to the 'word number form' described in section IV.2. The programs described in sections IV.3, IV.4 and IV.5 can then be used to construct vectors, calculate similarity coefficients and perform clustering for this word pair representation of the documents.

The following sections describe the program which converts the data to word pair number form.

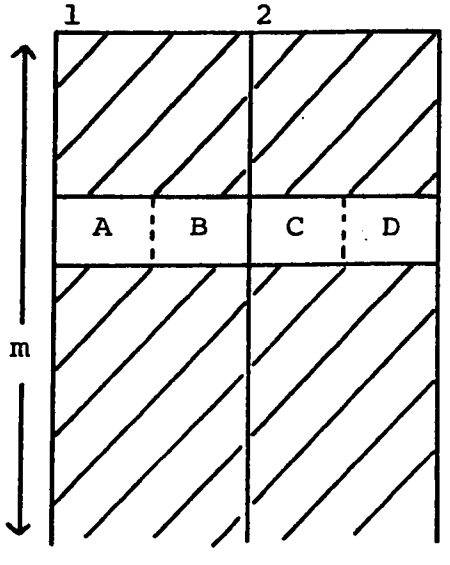
IV.6.1 Input and output

The program does not require the original text of the documents, but takes as input the single word number form of the text. Each distinct pair of adjacent word numbers input is assigned a distinct positive integer. There are no 'common' word pairs, but the common single words can be excluded from the pairs by ignoring any negative numbers input.

The program constructs a table of records for all word pairs identified. As each pair is found on input, the table is consulted to find the pair number if this pair has occurred before. If it has not, the pair is assigned the next positive integer and a new record is created in the table. The pair number, either retrieved from the table or newly assigned, is output to a file of the text in pair number form.

IV.6.2 Structure of the word pair table

The word pair table is a two-dimensional array representing a 2 x m matrix, where m is large enough to accommodate all distinct word pairs. A record in the table for a particular pair consists of the word numbers of the two words making the pair, their pair number and a count of the frequency of occurrence of the pair. The pair number and frequency count are packed in the first location of the record, and the two word numbers are packed in the second. The table is accessed by a hash address calculated from the two word numbers constituting the pair. The hash value for a given pair is the product of the two word numbers, normalised to lie within the range of the pair table. Figure IV.6 shows a typical entry in the word pair table.



A = Frequency of occurrence of pair
B = Word pair number
C = Number of first word of pair
D = Number of second word of pair

FIGURE IV.6 Record in word pair table

The file of the text in word pair number form is treated exactly the same as the text in single word number form to produce a single-link clustering of the document collection.

IV.7 Concordance

As a by-product of the output from the first two programs in the classification suite, we can construct a concordance of the words in the document collection, that is, an alphabetical index of the distinct words together with references to all the documents in which each word occurs.

IV.7.1 Construction of the concordance

We scan the word number arrays of all document vectors, looking for each word number in turn. Beginning with word number 1, we examine the first word number in each array, and if this is number 1 we store a reference to the document, and move on to the next location in the array. If the number is not found in the current location examined we know that it is not in the array at all, since the array is numerically ordered. In this case we remain at the current location for the next step.

For each subsequent word number we examine the locations currently reached in the arrays and repeat the above examination.

Common words are not normally concorded as they occur in nearly all documents. Hence, the vectors which exclude the common words are used for the above process.

IV.7.2 Structure of the concordance

The concordance is divided into blocks consisting of variable length records, one for each word number. The first location of each record contains the word number, and the second the number of references for that word. The document numbers follow in numerical order. For example, Figure IV.7(a) shows the concordance record for word 3 which occurs in eight documents altogether, including documents 1, 2 and 5.

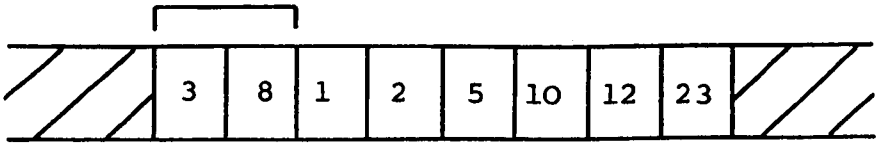


FIGURE IV.7(a) Typical concordance record

As each block of the concordance becomes full it is transferred to external storage. A two-dimensional array stores the number of the block in which a particular record is placed and the starting address of the record within that block. This 'block pointer array' is accessed by the word numbers; The pointers for word i are found in locations (i,1) and (i,2) of the pointer array. Figure IV.7(b) shows the relation between the pointer array and concordance blocks.

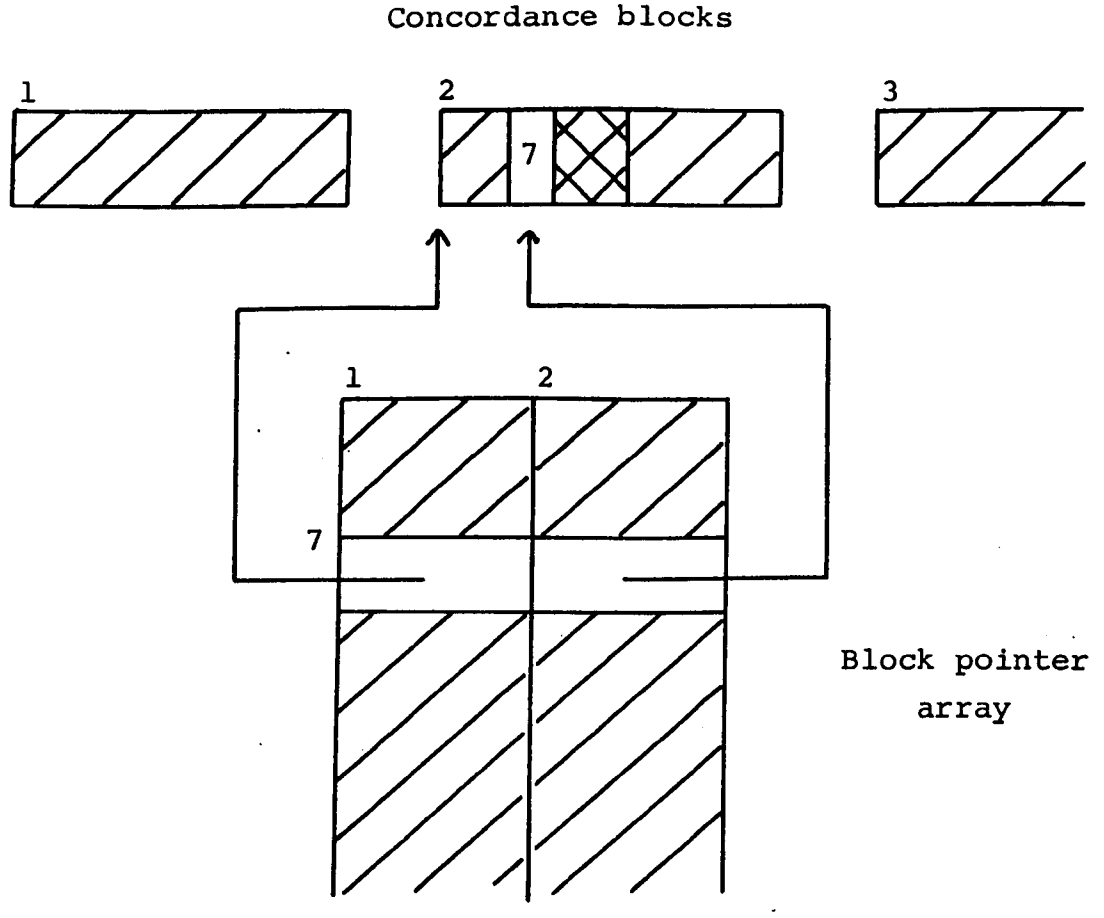


FIGURE IV.7(b) Structure of concordance

IV.7.3 Printing the concordance

The word numbers, in alphabetical order of the words they represent, are available in a file constructed by the dictionary program. These word numbers are processed in this order, to produce the alphabetical concordance. For each word number encountered, the block pointer array is consulted for the address of the concordance record. This record is read from the appropriate block, and the information contained in it is printed out with the word to which it refers. Common words are included in the list, but are followed by the characters '****' to indicate that no document references are given.

IV.8 Conclusion

A suite of programs has been described for generating a single-link clustering of a collection of documents, based on the frequency of occurrence of the words in the documents. A further program is included which converts the text to a suitable form for clustering on the basis of the frequency of occurrence of pairs of words. A concordance of the words in the document collection can be obtained as a by-product of the classification process, and a program is described for producing this information.

The programs have been implemented on two different computer installations, and have been used to classify the Council of Europe Conventions and Agreements and other smaller data bases. The results are presented and discussed in the following chapters.

CHAPTER V

CLASSIFICATION OF THE CONVENTIONS AND AGREEMENTS OF THE COUNCIL OF EUROPE

In this chapter we discuss the results of applying the theories of Chapters II and III to a large body of text, through implementation of the computer programs described in Chapter IV.

V.1 Data

Two data bases were used for the classification experiments. These consisted of the full text of the Agreements, Conventions and Protocols concluded between Member States of the Council of Europe, one version being in English, the other in French. This data was available in machine-readable form, held on magnetic tape.

There are 86 of these treaties, containing some 280,000 words, giving an average of 3,200 words per treaty. Each treaty is divided into articles and there are a total of 2,550 of these, each containing 110 words on average. The number of distinct English words in the treaties is 6,610, and for the French version the corresponding number is 8,305. Further statistics of this data are presented in Chapter VI.

The treaties are a self-contained set of documents which deal with a wide variety of topics including economics, social and cultural activities, and legal and administrative matters. Both the English and French versions are equally authentic, and considerable care has been taken by the draftsmen to ensure that expressions used in one are, as nearly as possible, equivalent to expressions used in the other. This allows us to make direct comparisons between the classifications produced for each of the two versions of the text.

The titles of the treaties are listed in Appendix B, together with the document numbers which are used as reference numbers by the computer programs. The number in brackets following each title is that assigned by the Council of Europe as a reference number. Some of the treaties have been split up for the data base, and so some of the titles listed have the same 'Council of Europe number'. Also some of the original treaties have been amalgamated for the data base.

V.2 Common words

The words which were to be regarded as 'common words' were chosen manually, on the basis of their frequencies of occurrence and distribution in the documents. A preliminary dictionary of words in the English text was compiled,

giving the total frequency of occurrence of each word and the number of different treaties in which it occurred. The usual common words, e.g. articles, prepositions and conjunctions, were found to be the most frequently occurring, and most of these occurred in over 80 treaties. These were selected as common words, and the list supplemented with those words which appeared to be common for this particular data base.

The extra common words chosen were those which occurred in 60 or more treaties, with a total frequency of 1,000 or more. Amongst these were found some single letters which, apart from the words 'A' and 'I', are used only to number sections of articles. Thus these words are merely functional and have no semantic content. All single letters were therefore input as common words, although some of these occurred infrequently. Other than these single letters, 44 English common words were selected, and these are listed in Table V.2(a) in order of frequency of occurrence.

A similar list of French common words was compiled by translating the English list. There are more French common words owing to the several gender and plural forms corresponding to one English word. A few extra words were also added which appeared to be common in the French text. The single letters were again taken to be common, and in fact several of these are meaningful words in French. The 60 meaningful French common words are listed in Table V.2(b).

Table V.2(a) Common words for English text

THE	26322	IS	1710
OF	21311	MAY	1633
TO	7942	STATE	1593
IN	7247	COUNCIL	1445
AND	5009	NOT	1378
OR	4693	IT	1362
A	4678	PARAGRAPH	1249
SHALL	4642	SUCH	1206
ARTICLE	4084	EUROPE	1204
BE	3304	PROVISIONS	1179
FOR	2837	ITS	1094
BY	2792	AGREEMENT	1092
THIS	2422	OTHER	1049
ANY	2268	IF	1031
WHICH	2061	AT	990
CONVENTION	1982	UNDER	907
CONTRACTING	1934	ARE	795
ON	1824	AN	783
WITH	1806	INTO	670
PARTY	1754	WHERE	641
THAT	1751	ALL	476
AS	1746	I	398

Table V.2(b) Common words for French text

DE	17670	EUROPE	1227
L	10999	NE	1186
LA	10174	IL	1138
A	8521	PRESENTE	1095
DES	5920	S	1052
D	5879	PAS	1049
DU	5868	ETRE	1029
LES	5854	TOUTE	1028
LE	5796	PARTIES	1015
ET	4849	PRESENT	1001
OU	4636	CONTRACTANTE	995
EN	4345	SONT	951
ARTICLE	4053	ACCORD	923
UNE	2741	TOUT	894
AU	2705	QU	874
PAR	2660	SI	864
UN	2268	CE	854
DANS	2207	CETTE	838
CONVENTION	2059	AUTRE	743
QUI	1968	N	634
AUX	1943	CES	576
POUR	1905	AUTRES	564
EST	1727	C	502
PARTIE	1717	SOUS	406

Table V.2(b) (continued)

QUE	1658	Y	397
ETAT	1611	CET	383
SUR	1524	AVEC	295
CONSEIL	1474	TOUS	285
DISPOSITIONS	1398	TOUTES	225
PARAGRAPHE	1231	J	41

In sections V.3, V.4 and V.5 following, classification is based on single words. That is, elements in the vectors which represent documents are the frequencies of occurrence of the individual words in the documents. Classification based on word pairs is briefly investigated in section V.6.

V.3 Classification of treaties

The first experiment was to consider each complete treaty as one document, and to cluster these on the basis of their word content. To begin, new dictionaries were constructed for the two texts with the common words, given in section V.2, identified by negative word numbers. The texts were converted to word number form, and this data used to construct document vectors both including and excluding the common words.

V.3.1 Common words excluded

To exclude the common words, just the positive word numbers were used in the document vectors. Coefficients of similarity between pairs of vectors were calculated, and were all non-zero. That is, each pair of treaties contained some identical non-common words. The range of coefficients for the two data bases is shown in Table V.3(a), and the level at which clustering was complete, that is where all documents belong to a single cluster, is also given.

Table V.3(a) Range of similarity coefficients for treaties with common words excluded

	English text	French text
Highest coefficient and level at which clustering begins	0.951	0.946
Level at which clustering is complete	0.066	0.075
Lowest coefficient	0.001	0.001

The documents were clustered according to these similarity coefficients, and the resulting dendrograms are shown in Figures V.3(A) and V.3(B) for the English and French texts respectively. The clustering is markedly similar in both cases, as might be expected because of the equivalence of the two texts.

At various levels of the similarity coefficient we are able to pick out some well-formed clusters of documents which, from their titles, appear to be semantically related. These clusters can be given meaningful headings, and these headings, which correspond to the labels attached to the clusters in Figure V.3(A), are listed in Table V.3(b). The total number of treaties which belong to each of these clusters is also given. The headings for the French clusters in Figure V.3(B) are also those listed in Table V.3(b). Where a French cluster differs slightly from the equivalent English one, then the label assigned to it is modified by adding the subscript 1. For example, the 'Human rights - protocol' cluster, cluster C in the English case, is labelled C_1 in the French dendrogram as it contains an additional treaty.

DENDROGRAM SHOWING

CLASSIFICATION OF

TREATIES.

ENGLISH TEXT, EXCLUDING

THE COMMON WORDS.

FIGURE V. 3 (A)

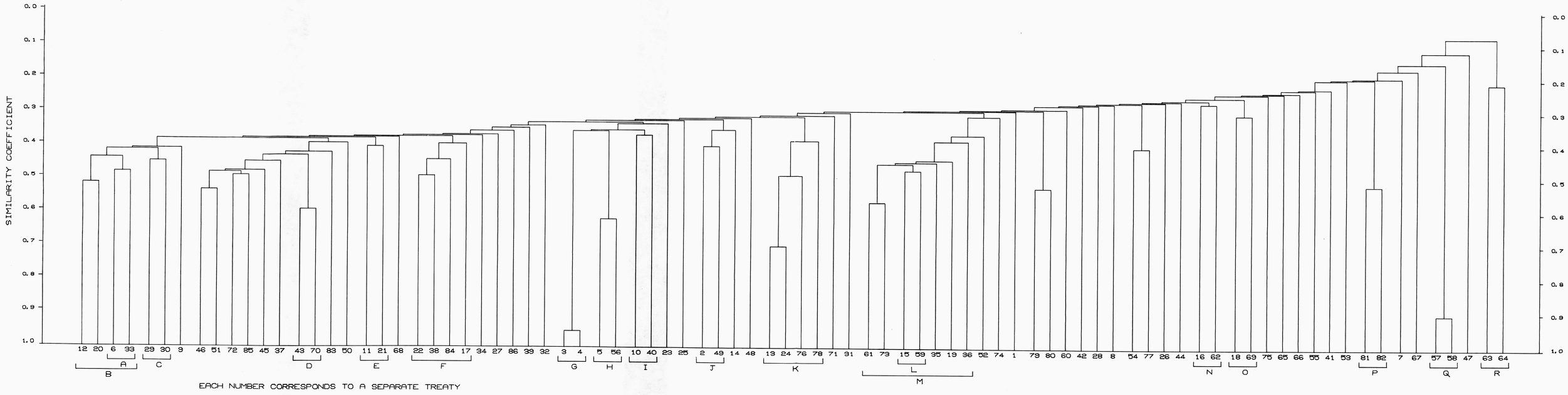

SIMILARITY COEFFICIENT
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
12 20 6 33 29 30 9 46 51 72 85 45 37 43 70 83 50 11 21 68 22 38 84 17 34 27 86 39 32 3 4 5 56 10 40 23 25 2 49 14 48 13 24 76 78 71 31 61 73 15 59 35 19 36 52 74 1 79 80 60 42 28 8 54 77 26 44 16 62 18 69 75 65 66 55 41 53 81 82 7 67 57 58 47 63 64
A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
P
Q
R
EACH NUMBER CORRESPONDS TO A SEPARATE TREATY

Table V.3(b) Headings for clusters of treaties

A	Equivalence of diplomas for entrance to university (2)
B	University study (4)
C	Human rights - protocol (2)
D	Money liabilities (2)
E	Medical treatment (2)
F	Television broadcasts (4)
G	Social security schemes (2)
H	Social and medical assistance (2)
I	Establishment (2)
J	Human rights (2)
K	Privileges and immunities (4)
L	Extradition (2)
M	Criminal matters (7)
N	Travel between member states (2)
O	Civil liability with respect to motor vehicles (2)
P	Social security (2)
Q	Annexes to social security schemes (2)
R	Exchange of therapeutic substances (2)

DENDROGRAM SHOWING

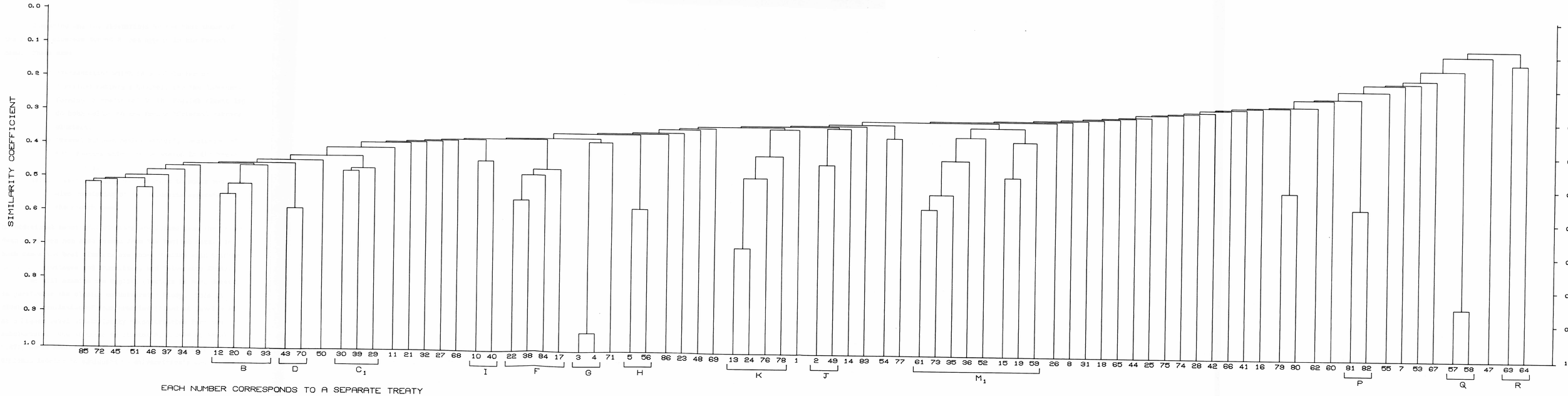
CLASSIFICATION OF

TREATIES.

FRENCH TEXT, EXCLUDING

THE COMMON WORDS.

FIGURE V. 3 (B)


SIMILARITY COEFFICIENT
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
85 72 45 51 46 37 34 9 12 20 6 33 43 70 50 30 39 29 11 21 32 27 68 10 40 22 38 84 17 3 4 71 5 56 86 23 48 69 13 24 76 78 1 2 49 14 83 54 77 61 73 35 36 52 15 19 59 26 8 31 18 65 44 25 75 74 28 42 66 41 16 79 80 62 60 81 82 55 7 53 67 57 58 47 63 64
B
D
C1
I
F
G
H
K
J
M1
P
Q
R
EACH NUMBER CORRESPONDS TO A SEPARATE TREATY

Comparing the two clusterings we see that three of the English clusters formed do not appear in the French case. These are:

(i) 'Extradition' which is a subcluster of 'Criminal matters'; however, the two documents forming 'Extradition' in the English clustering do both belong to the French 'Criminal matters' clusters.

(ii) 'Travel between member states', consisting of two treaties which are separated in the French clustering.

(iii) 'Civil liability with respect to motor vehicles', also consisting of two documents separated in the French case.

No additional meaningful clusters are formed from the French data which are not also formed from the English data. In both cases the best clusters are those dealing with Criminal Matters, Privileges and Immunities, and University Study.

We will examine the English clusters in more detail, in particular the cluster on University Study. This is cluster B consisting of four treaties joined at level 0.435. At a higher level of similarity these treaties form two subclusters. One of these is cluster A, formed at level 0.477, whose member treaties deal with the 'equivalence of diplomas leading to admission to universities'; one treaty

is the Convention on this topic, the other is the Protocol to this Convention. Thus we have a homogeneous pair of treaties.

The other subcluster, formed at level 0.510, does not appear so homogeneous from the titles of the treaties alone, and has therefore not been given a heading. One of these treaties deals with 'the equivalence of periods of university study', document 12, and the other deals with 'the recognition of university qualifications', document 20. However, examining the text reveals that 12 recognises a period of study at a foreign university as equivalent to one spent at a home university, provided the appropriate examinations are passed at the end of that period. That is, it recognises the qualifications obtained from a foreign university, which is similar in effect to document 20. The main difference is that 12 recognises a qualification from a foreign university as contributing to a particular qualification issued by a home university, whereas 20 recognises foreign qualifications as satisfying pre-requisites for courses to be followed at a home university.

The University Study cluster is joined by cluster C, Protocols to the Convention on Human Rights, as the level of the similarity coefficient decreases to 0.411. However, the two treaties comprising C do not mention university study at all. They are very short documents consisting mainly of

formal clauses of signature, accession, ratification etc. These formal clauses occur in all treaties, and are probably responsible for joining cluster C to cluster B.

The joint cluster B and C is joined by document 9 at level 0.407. This is the European Cultural Convention, which aims to promote the study of the language, history and civilisation of the member states, and so has clustered with the other 'study' treaties.

From its title it would seem that document 51 should also have joined the 'study' cluster, as it deals with the 'payment of scholarships for study abroad'. However, it has clustered with document 46, the agreement restricting the use of detergents. It is again likely that the clustering is due to the formal clauses, as otherwise the contents of these two documents are completely unrelated.

Another well-formed homogeneous cluster is cluster M on 'criminal matters'. The cluster label has not included document 52, whose title merely states that it is about the 'repatriation of minors'. However, this document clusters quite strongly with those on criminal matters, and on reading the text we find that it deals in part with the procedure to be adopted when a minor who is to be repatriated is the subject of criminal prosecution.

Both the English and French clusterings exhibit a large amount of chaining, a feature of single-link clustering discussed in section III.4.4. Though not particularly

useful in itself, this chaining is probably a true representation of the structure of the data contained in the treaties. There are not many compact, homogeneous groups of treaties; the treaties are a collection of individuals which mostly deal with different topics. As a result, the clustering is straggly and appears arbitrary in places, due to the effect of the formal clauses as stated earlier. The behaviour of the clauses taken individually as documents is examined in section V.4.

V.3.2 Common words included

To examine the influence of common words on the clustering, vectors were constructed for the complete text of the treaties, that is, both positive and negative word numbers were included. The similarity coefficients calculated for pairs of these vectors proved to be much greater than those for vectors which excluded the common words. The high frequencies of occurrence of the common words dominate the other terms in the coefficient and give a value which is closer to the maximum value of 1. The range of the coefficients in this case is given in Table V.3(c).

Table V.3(c) Range of similarity coefficients for treaties with common words included

	English text	French text
Highest coefficient and level at which clustering begins	0.997	0.994
Level at which clustering is complete	0.564	0.620
Lowest coefficient	0.352	0.379

The dendrogram representation of the clusterings of the vectors with common words included are shown in Figures V.3(C) and V.3(D). These dendrograms are very similar in parts to those for the clustering with common words excluded. Despite the change in absolute value of the coefficients, their order has been preserved on the whole, and it is this order which determines the clustering.

The headings in Table V.3(b) again apply to these dendrograms, with appropriate modifications made to the labelling where the clusters differ slightly from those in Figure V.3(A). (The labels for the English clusters are subscripted by 2, and those for the French subscripted by 3 where necessary.)

DENDROGRAM SHOWING

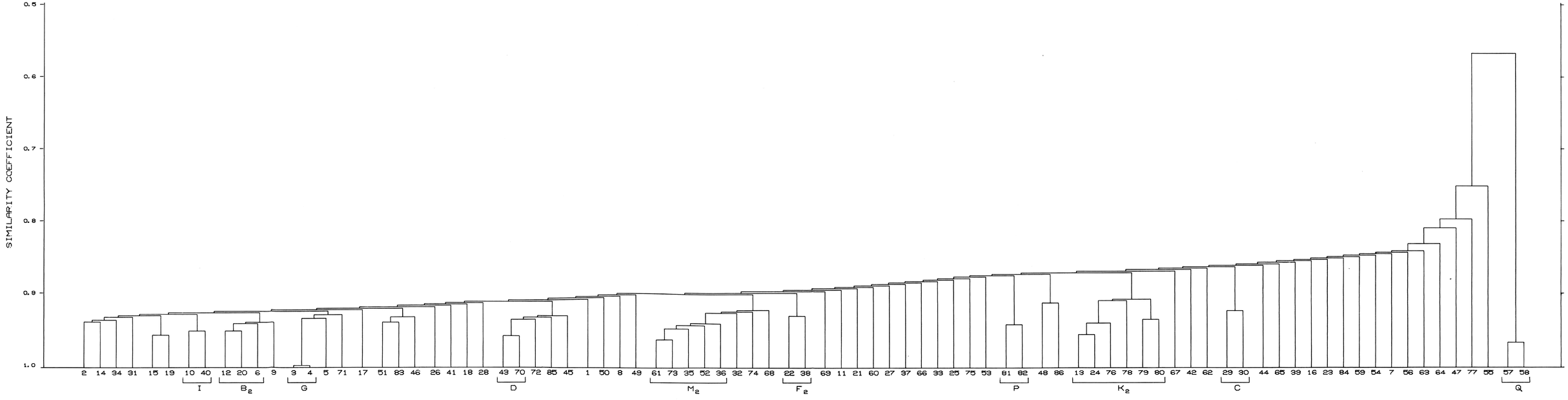
CLASSIFICATION OF

TREATIES.

ENGLISH TEXT, INCLUDING

THE COMMON WORDS.

FIGURE V. 3 (C)


SIMILARITY COEFFICIENT
0.5
0.6
0.7
0.8
0.9
1.0
2 14 34 31 15 19 10 40 12 20 6 9 3 4 5 71 17 51 83 46 26 41 18 28 43 70 72 85 45 1 50 8 49 61 73 35 52 36 32 74 68 22 38 69 11 21 60 27 37 66 33 25 75 53 81 82 48 86 13 24 76 78 79 80 67 42 62 29 30 44 65 39 16 23 84 59 54 7 56 63 64 47 77 55 57 58
I
B_2
G
D
M_2
F_2
P
K_2
C
Q
EACH NUMBER CORRESPONDS TO A SEPARATE TREATY

DENDROGRAM SHOWING

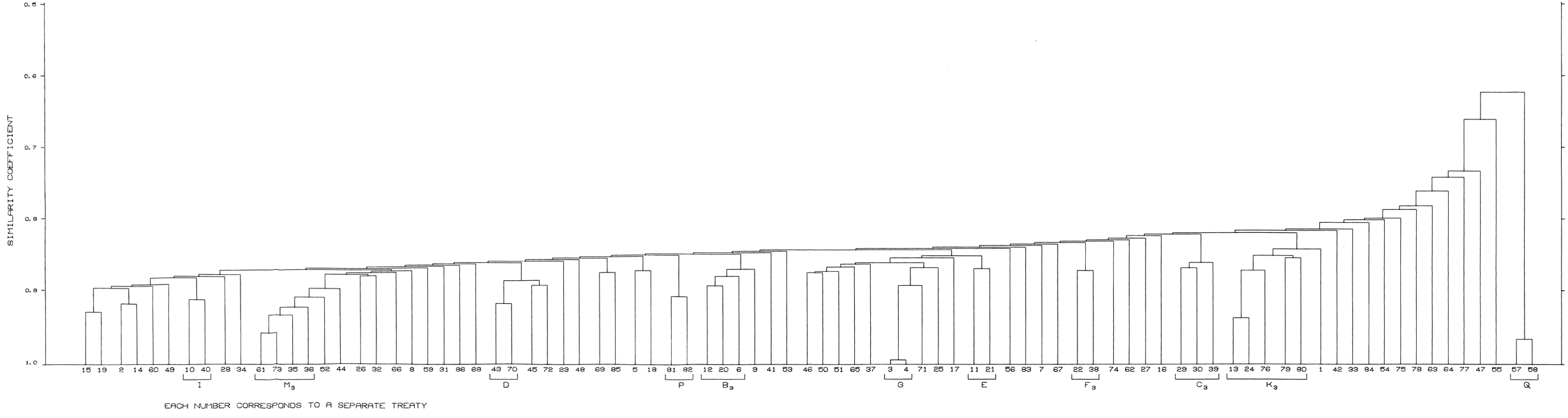
CLASSIFICATION OF

TREATIES.

FRENCH TEXT, INCLUDING

THE COMMON WORDS.

FIGURE V.3 (D)


SIMILARITY COEFFICIENT
0.5
0.6
0.7
0.8
0.9
1.0
15 19 2 14 60 49 10 40 28 34 61 73 35 36 52 44 26 32 66 8 59 31 86 88 43 70 45 72 23 48 69 85 5 18 81 82 12 20 6 9 41 53 46 50 51 65 37 3 4 71 25 17 11 21 56 83 7 67 22 38 74 62 27 16 29 30 39 13 24 76 79 80 1 42 33 84 54 75 78 63 64 77 47 55 57 58
I
M3
D
P
B3
G
E
F3
C3
K3
Q
EACH NUMBER CORRESPONDS TO A SEPARATE TREATY

The English clustering produced 10 of the 18 clusters which were formed originally from the English text with common words excluded. These same 10 clusters, with a few variations, plus one other cluster were formed in the French case. Since the French text with common words excluded produced only 13 well-defined clusters, the change in clustering is less marked than in the English case. No new meaningful clusters are found for either text in this case, and overall the clustering appears better when the common words are excluded.

V.3.3 Comparison with manual classification

The Council of Europe uses its own manual classification for arranging the contents of its publications, in reference (56) for example. Table V.3(d) lists the headings of the classes used, and gives the reference numbers of the documents assigned to each class. (These are the document numbers used by this study, not those used in the Council of Europe numbering system.)

Note that documents 1, 54, 55, 77 and 86 are not included in this classification.

Table V.3(d) Council of Europe Classification Scheme

Privileges and Immunities	13, 24, 76, 78, 79, 80
Human Rights	2, 29, 30, 31, 39, 49
Social Matters	3, 4, 5, 23, 41, 50, 56, 57, 58, 67, 81, 82
Public Health	11, 21, 25, 42, 63, 64, 65, 83
Public Health (partial agreement)	34, 46
Cultural Matters	6, 9, 12, 20, 33, 48, 51
Patents	7, 8, 32
Television	17, 22, 37, 38, 84
Public International Law	14, 66, 74, 75
Other Legal Subjects	10, 15, 18, 19, 26, 27, 28, 35, 36, 40, 43, 44, 45, 47, 52, 53, 59, 60, 61, 68, 69, 70, 72, 73, 85
Movement of Persons	16, 62, 71

Comparing the classes of the Council of Europe scheme with the clusters in Figure V.3(A) we find that the classes 'Privileges and Immunities', 'Cultural Matters' and 'Television' are very similar to the clusters 'Privileges and Immunities', 'University Study' and 'Television Broadcasts'. The documents forming the clusters 'Human Rights' and 'Human Rights - Protocol' are all members of the class 'Human Rights', and the clusters 'Social Security', 'Social Security Schemes', 'Annexes to Social Security Schemes' and 'Social and Medical Assistance' are joined together in the manual scheme to form part of the class 'Social Matters'.

The cluster 'Travel between Member States' is a subset of the class 'Movement of Persons', which also contains the treaty on the 'abolition of visas for refugees'. The large class 'Other Legal Subjects' is broken down by clustering into smaller, more specific groups. These clusters are 'Establishment', 'Extradition', 'Money Liabilities', 'Civil Liability with Respect to Motor Vehicles' and 'Criminal Matters'.

In general we can say that documents belonging to the same cluster are assigned to the same class by the manual scheme, though the reverse is not always true. The manual classification is a more general scheme than the clustering. It has aimed at being complete in the sense that no document is left in a class on its own, but each is assigned to the

most appropriate of the pre-specified classes. This strategy has produced the large class 'Other Legal Subjects' which contains those documents which do not fit neatly into any other category. Clustering, on the other hand, reveals more specific relations between documents, though at any one level in the dendrogram not all documents are necessarily assigned to a cluster.

The homogeneous clusters formed at a high level of similarity are subsets of the manually produced classes, but the two classifications differ widely at a lower clustering level. The manual classification is based on the main topic of each treaty. For example, document 71 joins the class 'Movement of Persons' because it is about visas used for travel. In contrast, the full-text, automatic method takes into account all the information in each document, including that contained in the formal clauses. Thus at a general level the documents are clustered on the basis of their dealing with the signature, ratification etc., of Council of Europe treaties.

Because of the behaviour of the formal clauses, the clustering is not very useful at a low level of similarity. Of course, if the treaties formed part of a larger collection we might wish to group them together as Council of Europe documents. It would be interesting to cluster some of the treaties together with documents from another source to see if they cluster by origin or topic.

In most legal information retrieval systems a 'document' is not as large as a whole treaty, but usually consists of one section or article of an act or treaty. In this case the formal clauses should not be a problem, as we would expect them to form clusters amongst themselves, and the articles containing the substance of the treaties to cluster on a subject basis. To test this idea the second clustering experiment was carried out and is described in the following sections.

V.4 Classification of articles of treaties

In this section and section V.5 the word 'document' means an article of a treaty. It would be unrealistic to attempt to classify, for test purposes, all the 2,550 articles as this requires the calculation of over 3 million similarity coefficients, and the resulting dendrogram would be too elaborate to represent and examine easily. For these reasons a small subset was selected, consisting of the four treaties in the 'University Study' cluster, and the 49 articles comprising these treaties were clustered. This homogeneous set of treaties was chosen so that we could expect to get some clustering of the non-formal clauses as well as the formal clauses. The English text of the 49 articles is given in Appendix C, with the document numbers used to reference the articles in the computer programs and the resulting dendrograms.

Both the English and French versions of these articles were clustered. The common words were excluded, as the first experiment with treaties gave better results for this case.

The text of the articles was already in word number form as a result of the first experiment, and this data was re-used for constructing vectors for the articles. This meant, of course, that the common words were the same as those used for the treaties, some of which were originally chosen for their high frequencies and even distribution in the complete collection of treaties. These conditions do not necessarily apply for a subset of the treaties, and in fact some of the content-bearing common words were not strictly common words in the particular set of articles used. However, as there were only a few such words this problem was ignored.

V.4.1 Article clusters

The usual range of similarity coefficients for pairs of articles is given in Table V.4.(a). Unlike the treaties, not all pairs of articles have a non-zero coefficient; that is, there are pairs which have no words the same. In fact 10.7% of the coefficients are zero for the English text, and 13.7% in the French case.

Table V.4(a) Range of similarity coefficients for articles with common words excluded

	English text	French text
Highest coefficient and level at which clustering begins	0.947	0.954
Level at which clustering is complete	0.061	0.036
Lowest coefficient	0.000	0.000

The dendrogram resulting from the clustering of the English text of the articles is shown in Figure V.4(A), and Figure V.4(B) is the French dendrogram. Again we are able to pick out some well-formed clusters at various levels of similarity. The headings which have been given to these clusters and which correspond to the labels in Figure V.4(A) are listed in Table V.4(b). Where the French clusters differ slightly, the labels in Figure V.4(B) are subscripted by 1, otherwise the same headings apply.

DENDROGRAM SHOWING

CLASSIFICATION OF

ARTICLES.

ENGLISH TEXT, EXCLUDING

THE COMMON WORDS.

FIGURE V. 4 (A)

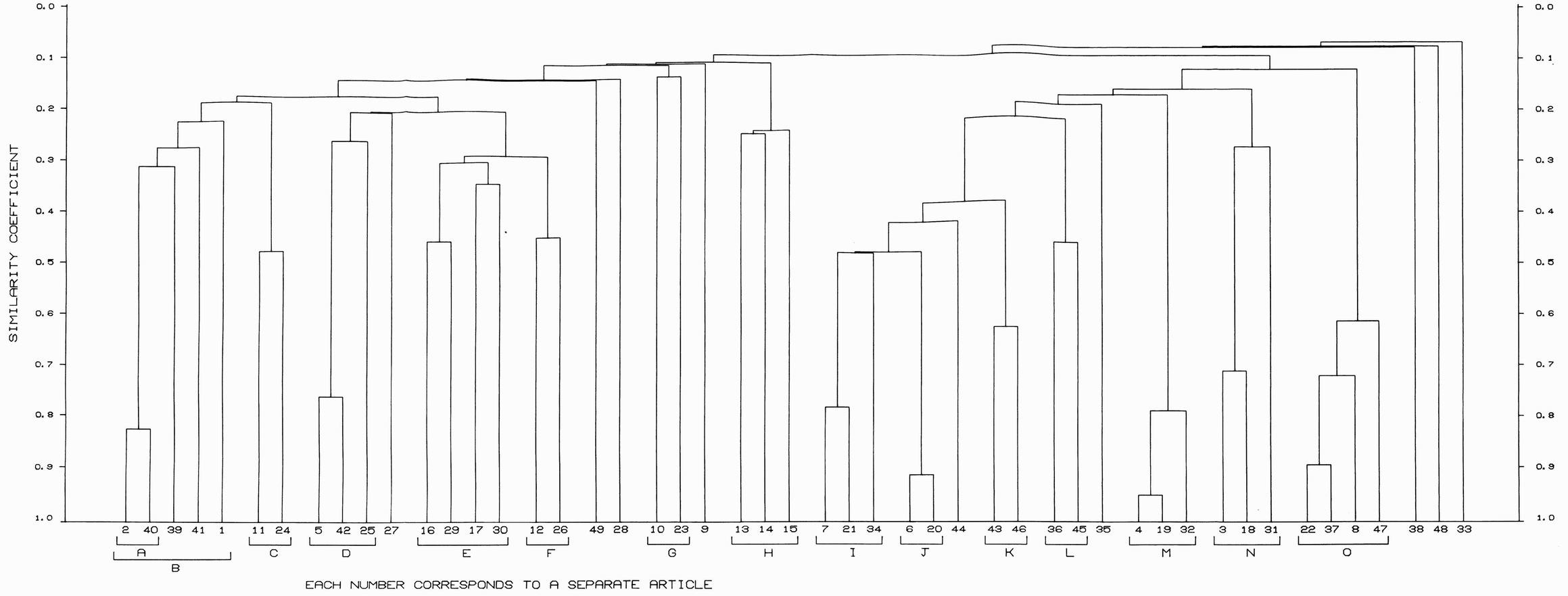

SIMILARITY COEFFICIENT
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
2 40 39 41 1
11 24
5 42 25 27
16 29 17 30
12 26
49 28
10 23 9
13 14 15
7 21 34
6 20 44
43 46
36 45 35
4 19 32
3 18 31
22 37 8 47
38 48 33
A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
EACH NUMBER CORRESPONDS TO A SEPARATE ARTICLE

Table V.4(b) Headings for clusters of articles

A	Agreements (2)
B	Equivalence of diplomas for entrance to university (5)
C	Preambles (2)
D	Definitions of terms used (3)
E	Application of provisions by education authorities (4)
F	Education authorities (2)
G	Declarations (2)
H	Equivalence of periods of study (3)
I	Accession by non-members of the Council (3)
J	Ratification (2)
K	Accession to protocol (2)
L	Denunciation of conventions (2)
M	Communication between Secretary-General and contracting parties (3)
N	Written statement of measures taken to implement conventions (3)
O	Signatures (4)

DENDROGRAM SHOWING

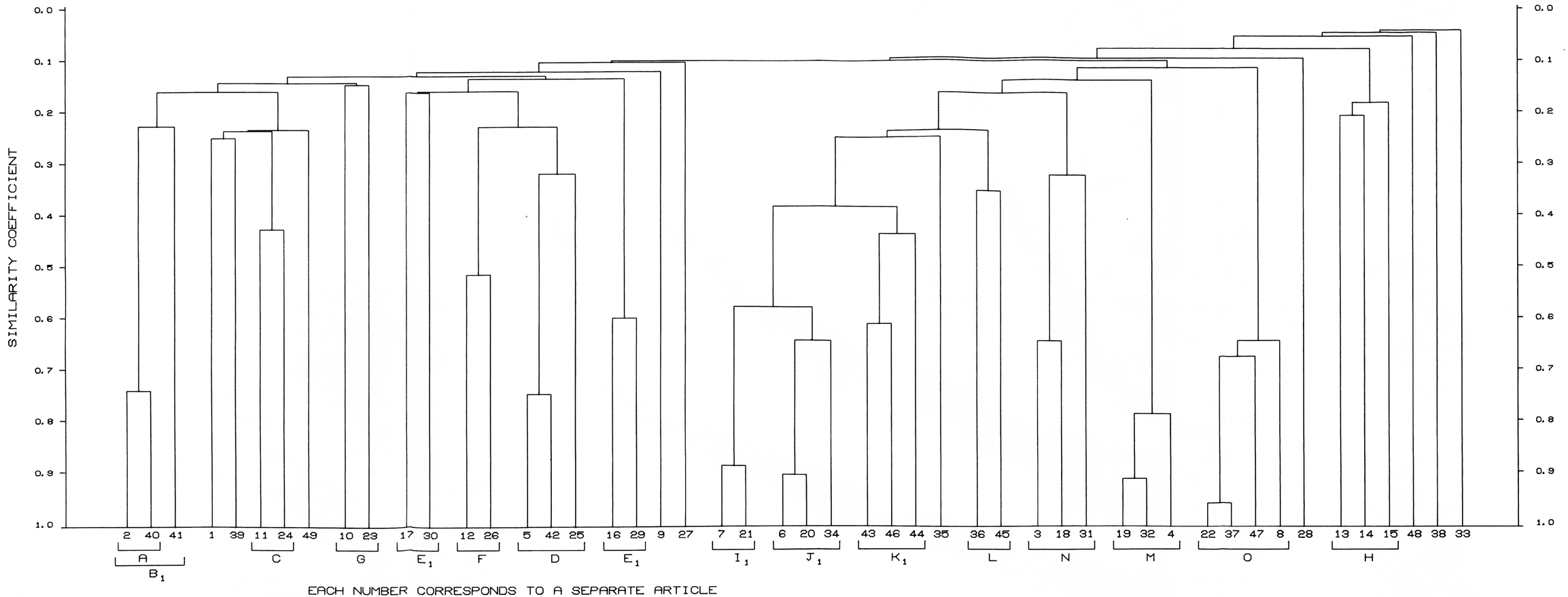
CLASSIFICATION OF

ARTICLES.

FRENCH TEXT, EXCLUDING

THE COMMON WORDS.

FIGURE V. 4 (B)


SIMILARITY COEFFICIENT
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
2 40 41
1 39 11 24 49
10 23
17 30
12 26
5 42 25
16 29
9 27
7 21
6 20 34
43 46 44 35
36 45
3 18 31
19 32 4
22 37 47 8 28
13 14 15 48 38 33
A
B1
C
G
E1
F
D
E1
I1
J1
K1
L
N
M
O
H
EACH NUMBER CORRESPONDS TO A SEPARATE ARTICLE

As expected the clustering is very similar for the two versions, English and French. More important, though, is the fact that the clauses have clustered in the way we hoped. We have formal clusters of 'Ratification', 'Signatures', 'Accession to protocol' etc., and also non-formal clusters, for example 'Equivalence of diplomas for entrance to university' and 'Definitions of terms used'. As these compact clusters merge at successively lower levels of similarity, we find that all the formal clusters merge into one cluster, and the non-formal ones form a separate cluster. The two join together at a very low level, 0.087 in the English case, and 0.096 in the French.

In the clustering of treaties many of the documents are not included in a labelled cluster at any level of similarity. However, owing to the choice of a homogeneous set of documents for article clustering, there are very few isolated documents in this case. We will examine more closely the contents of those articles which do not fit into any of the labelled clusters, in the English case.

Document 27 has clustered with the articles which define certain terms used in the treaties, because it uses these particular terms very frequently itself. However, the article is specifically about the recognition of academic qualifications, and therefore cannot be included in cluster D.

Of a more general nature are documents 28 and 49, which deal with the equivalence of diplomas and the passing of examinations. They are not highly similar to any of the more specific articles on these topics, and join the union of the relevant clusters at a relatively low level of similarity. Another isolated document is 9, a declaration for Belgium. Although this is a formal clause, it has clustered with the non-formal clauses as it specifically mentions the 'equivalence of diplomas leading to admission to universities'.

Two of the documents in the merged formal cluster are not included in any specific labelled cluster. One of these is document 44 which deals with the date when the Protocol to the Convention on the Equivalence of Diplomas shall come into force, relative to when the contracting party ratified or acceded to that Convention. 'Ratification' and 'accession' contribute almost equally to this document, and so it joins the union of the two clusters 'Ratification' and 'Accession of non-members'. The other isolated document is 35, which is not specifically related to any other article, but uses many of the general terms used in the formal clauses and has clustered with them arbitrarily.

Three documents join the clustering after the formal and non-formal clusters have merged. Two of these, 38 and 48, are declarations for the Kingdoms of Greece and of the Netherlands, and are not specifically related to the other

articles. The other, document 33, unfortunately contains a large number of the content-bearing common words, and uses these to discuss a point of general interest to the other articles. Because these common words were excluded, this document has a very low coefficient of similarity with all other documents and is the last to cluster.

V.5 Classification of sentences

As an interesting exercise, a subset of the articles used in the preceding section was selected, and broken down into the constituent sentences to be clustered.

The sentences used were those comprising documents 2, 39 and 40. Documents 2 and 40 form the cluster 'Agreements' in Figure V.4(A), and 39 joins them to form part of cluster B on the 'equivalence of diplomas'. These three documents are Article 1 of the Convention on the Equivalence of Diplomas, the Preamble for the Protocol to this Convention, and Article 1 of this Protocol. The text can be found in Appendix C.

The articles were divided into eleven sentences in all, numbered 1 to 4 for document 2, 5 to 7 for document 39 and 8 to 11 for document 40. The English text only was used, with the common words excluded. Table V.5(a) gives the range of similarity coefficients, and Figure V.5 shows the resulting dendrogram of clustered sentences. Headings for the labelled clusters are listed in Table V.5(b).

Table V.5(a) Range of similarity coefficients for sentences with common words excluded

	English text
Highest coefficient and level at which clustering begins	1.000
Level at which clustering is complete	0.022
Lowest coefficient	0.000

Table V.5(b) Headings for clusters of sentences

A Admission subject to availability of places (2)

B Informing universities of Convention (2)

C Equivalence of diplomas (4)

D Right not to apply Convention (2)

DENDROGRAM SHOWING

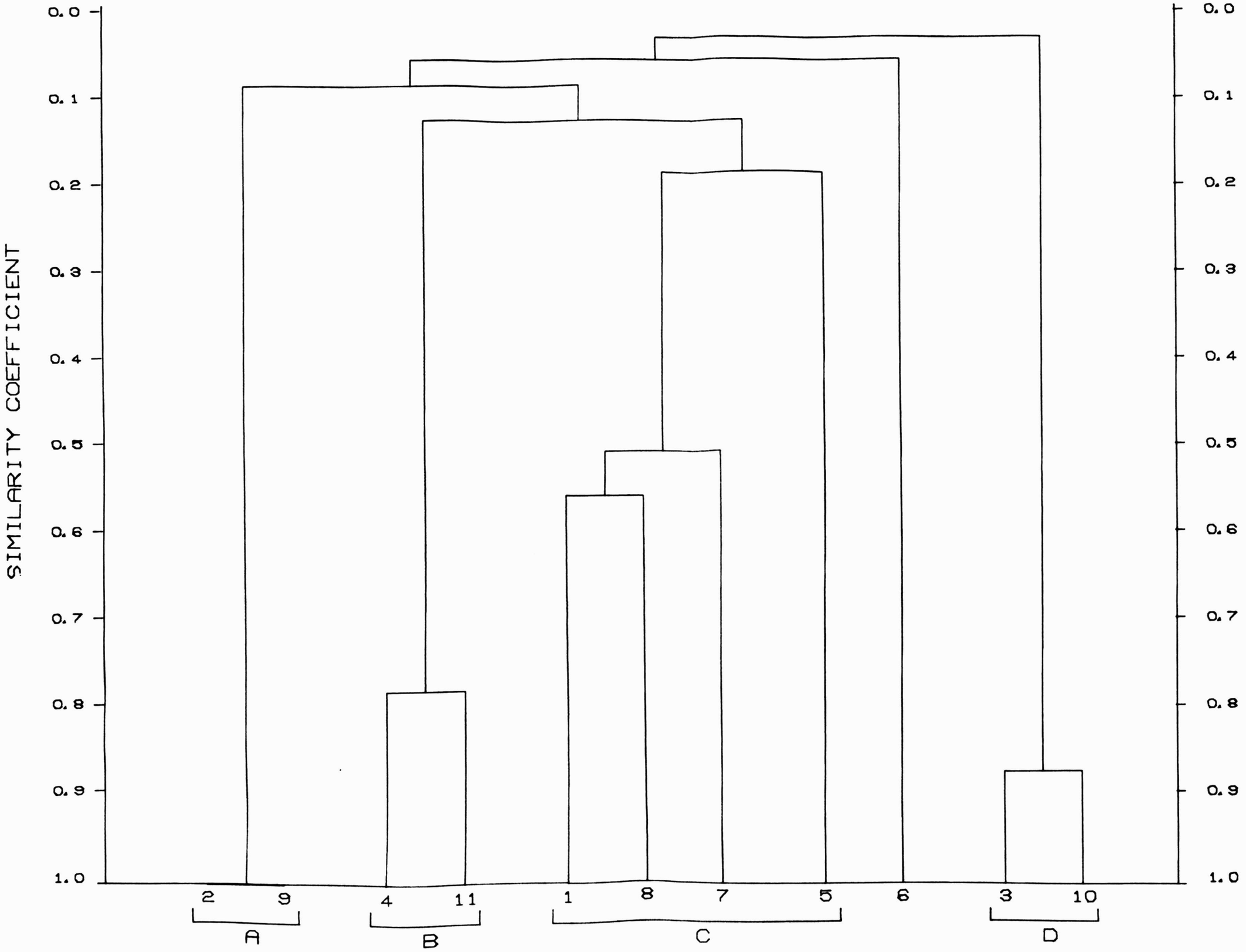
CLASSIFICATION OF

SENTENCES.

ENGLISH TEXT, EXCLUDING

THE COMMON WORDS.

FIGURE V.5


SIMILARITY COEFFICIENT
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
2 9
4 11
1 8 7 5
6
3 10
A
B
C
D
EACH NUMBER CORRESPONDS TO A SEPARATE SENTENCE

From Figure V.5 we see that the pair 2 and 9 has the maximum coefficient of similarity 1. That is, their word content apart from common words must be identical, and in fact they are identical overall.

Eight of the sentences have formed four clusters at a high level, above 0.56. Examining the two articles which these eight sentences constitute, we see that they are very similar in style. Each contains four sentences, and these deal with similar topics in the same order within the article. That is, sentence 1 of one article is very similar to sentence 1 in the other, and so on. In each case the first sentence is the main content-bearing one. Each of the sentences is more similar to the corresponding one in the other article than to any of the other sentences in the same article.

At a lower level of similarity the sentences from the Preamble join the cluster consisting of 1 and 8, to form cluster C which contains the substance of the articles. The pair 3 and 10 is last to join the merged cluster. These two sentences are of a more formal nature, dealing with the right not to apply the Convention.

The analysis of sentences indicates why particular articles cluster together. For example, it is likely that sentence 7, the third sentence of document 39, was responsible for this document joining the article cluster 'Agreements', as this sentence is the most similar to the first sentence

of each of the documents 2 and 40. The other sentences in 39 are not very similar to any in 2 and 40.

V.6 Classification based on word pairs

In this section we discuss the use of pairs of words as a basis for classifying the treaties. The vector representations of the documents to be classified contain the frequency of occurrence of the word pairs in the documents.

The text was converted to word pair number form using the program described in section IV.6. Two versions were created for each language, one including the common words, the other excluding them. A document in this case was taken to be a whole treaty. The clusterings obtained were not as good as those based on single words, and so not all the results are presented.

Table V.6 gives the range of similarity coefficients for the case where the common words were excluded. Approximately 3% of the pairs of English treaties have no word pairs the same, that is, have zero coefficient of similarity. The corresponding number in the French case is 2.4%. In general, the coefficients are much lower than when single words were used.

Table V.6 Range of similarity coefficients for treaties with common words excluded, using word pairs

	English text	French text
Highest coefficient and level at which clustering begins	0.850	0.830
Level at which clustering is complete	0.021	0.016
Lowest coefficient	0.000	0.000

The dendrogram for the English clustering, with common words excluded, is shown in Figure V.6. The headings in Table V.3(b) apply to these clusters, with the labels subscripted by 4 where the clusters vary from those in Figure V.3(A). Fewer compact clusters are formed in this case, and much more chaining occurs in comparison with the single word clustering.

It is possible that word pairs describe the treaties too specifically, so that we get very little overlap between pairs of treaties in terms of word pair content. Classification summarises document descriptions, and for this purpose they need to be more general.

DENDROGRAM SHOWING

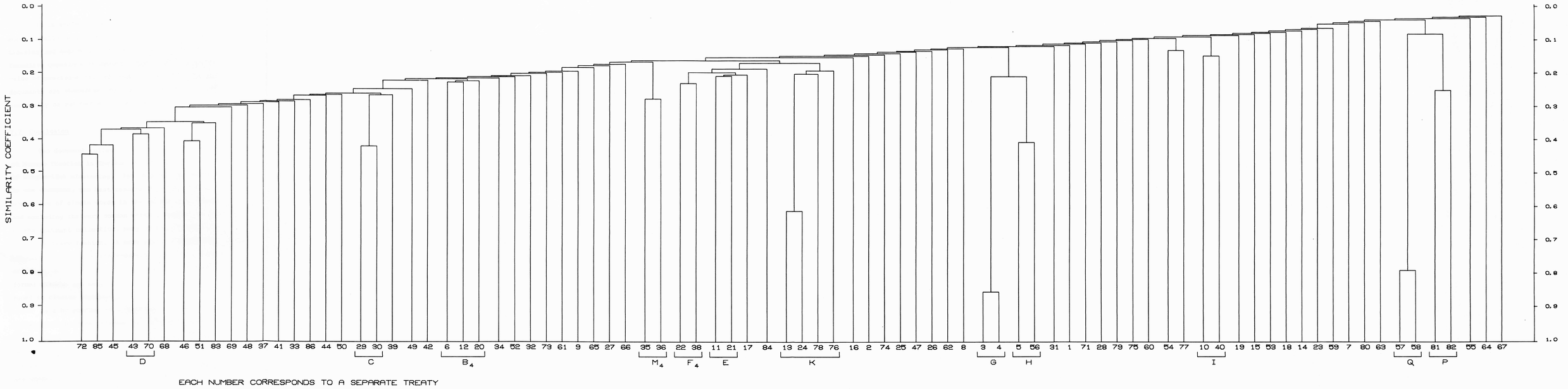
CLASSIFICATION OF

TREATIES.

ENGLISH TEXT. WORD PAIRS,

EXCLUDING COMMON WORDS.

FIGURE V.6


SIMILARITY COEFFICIENT
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
72 85 45 43 70 68 46 51 83 69 48 37 41 33 86 44 50 29 30 39 49 42 6 12 20 34 52 32 73 61 9 65 27 66 35 36 22 38 11 21 17 84 13 24 78 76 16 2 74 25 47 26 62 8 3 4 5 56 31 1 71 28 79 75 60 54 77 10 40 19 15 53 18 14 23 59 7 80 63 57 58 81 82 55 64 67
D
C
B4
M4
F4
E
K
G
H
I
Q
P
EACH NUMBER CORRESPONDS TO A SEPARATE TREATY

It is thought that the notion of word pairs could more usefully be incorporated into a retrieval system, for indexing and searching. In this case we need to be able to formulate questions in specific terms for a reasonable value of precision for the search. Specific descriptions of documents are therefore required so that the relevant ones can easily be matched with the questions.

V.7 Conclusion

Two document collections, consisting of the Council of Europe Treaties in English and French, have been classified by single-link clustering. Taking each complete treaty to be one document, the best results were obtained using the occurrence of single words in the vector representations, and excluding the very common words. Because of the nature of the document collection, many of the treaties clustered in an arbitrary fashion. A more useful classification was obtained by splitting the treaties into their constituent articles, and regarding each of these as a document. The formal clauses and subject-bearing, non-formal clauses were found to cluster separately.

As a by-product of processing the treaties for classification, some interesting statistics were obtained of the vocabulary used in these documents. Comparisons can be made between the English and French terminology in terms of variety of words and their frequency of use. These statistics are presented and discussed in the following chapter.

CHAPTER VI

STATISTICS OF THE CONVENTIONS AND AGREEMENTS OF THE COUNCIL OF EUROPE

VI.1 Introduction

The statistics of language, in particular written language, is of interest not only for its own sake but also for its numerous practical applications. It has been applied to the construction of code systems such as Morse Code and shorthand, to produce an economical use of symbols. It has also been used for describing literary style and for determining authorship, and Chapter VII discusses an example of this particular application. For information retrieval purposes, a statistical analysis of text can help to determine the choice of common words, index terms and weighting functions. These and many more uses are described in detail by Good (57), who also provides an extensive bibliography of studies in language statistics.

The data collected during computer processing of the Council of Europe Conventions and Agreements provides statistical information on the words and word pairs which were used as coordinates in the vector representation of these documents. Some of the more interesting results are presented in the following sections.

VI.2 Statistics of single words

VI.2.1 Analysis of number of words used

Table VI.2(a) gives the total number of words occurring in each of the English and French texts, and the number of these which are different. There are more different French words than there are English due mainly to the variety of gender forms for certain terms, e.g. tout, tous, toute, toutes, for which there is just one equivalent English word. The French language also uses more words than English to express the same idea, indicated by the greater total of words used for an equivalent set of documents. This is another reason for the greater number of different words.

Table VI.2(a) Number of words used in the treaties

	English text	French text
Total number of words	274,452	287,600
Number of different words	6,610	8,305

In addition to these totals for the whole data base, it was also possible to obtain cumulative sums of the number of words used in each treaty. Graphs were drawn from these values to show how the number of different words increases with the increase of the total number of words.

Figure VI.2(a) shows these graphs, plotted on a log-log scale. Both curves can be approximated to a straight line given by the equation:

$$\log W_D = A \log W_T + B$$

where W_D = the number of different words

W_T = the total number of words

and A and B are constants.

In the English case A and B are approximately 0.52 and 1.01 respectively, and are 0.52 and 1.11 respectively for the French graph.

The deviations form the straight line occur similarly in both cases, as expected. New topics, and hence new words, are introduced in corresponding documents in each data base. There is a particularly sharp rise in the number of different words after document 63 has been processed, at points (153613, 5200) and (160987, 6594) on the English and French graphs respectively. This document is the agreement on the exchange of therapeutic substances, an entirely new

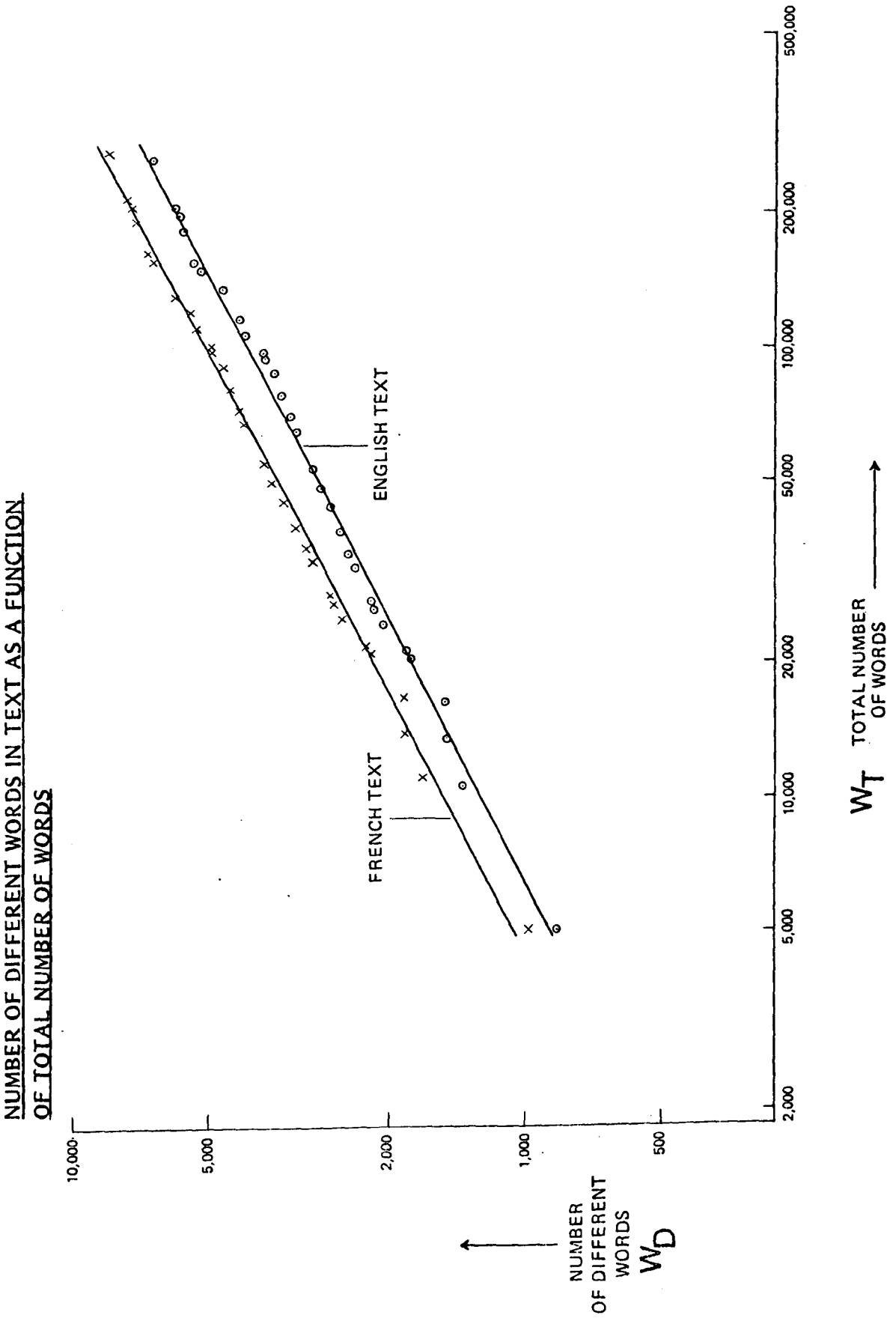

NUMBER OF DIFFERENT WORDS IN TEXT AS A FUNCTION OF TOTAL NUMBER OF WORDS
ENGLISH TEXT
FRENCH TEXT
NUMBER OF DIFFERENT WORDS W_D
TOTAL NUMBER OF WORDS W_T
10,000
5,000
2,000
1,000
500
2,000
5,000
10,000
20,000
50,000
100,000
200,000
500,000

FIGURE VI.2(a)

topic not discussed in any of the preceding documents. Thus it contains many new words which do not occur previously. We see from the dendrogram in Figure V.3(A) of the previous chapter that this treaty and the one following, document 64, are very dissimilar from the rest of the collection.

VI.2.2 The most frequent words

The 30 most frequent words in each of the two texts are precisely the 30 most frequent of the common words chosen in Chapter V and listed in Tables V.2(a) and V.2(b). Overall, the common words constitute approximately 50% of the text in each of the two versions, and these words are mainly responsible for the characteristics of the text discussed in the following section.

VI.2.3 Analysis of word lengths

One of the data items recorded for each word in the dictionary was the number of letters in the word, and it was therefore possible to analyse the distribution of word lengths in the text. Figures VI.2(b) and VI.2(c) show, for English and French respectively, the percentage of the total number of words having a particular number of letters.

The general shape of the distribution of word lengths is similar in both cases, though the absolute values of the percentages vary considerably in places. For example, words

HISTOGRAM SHOWING THE NUMBER OF CHARACTERS IN A WORD AS A FUNCTION OF THE TOTAL NUMBER OF WORDS

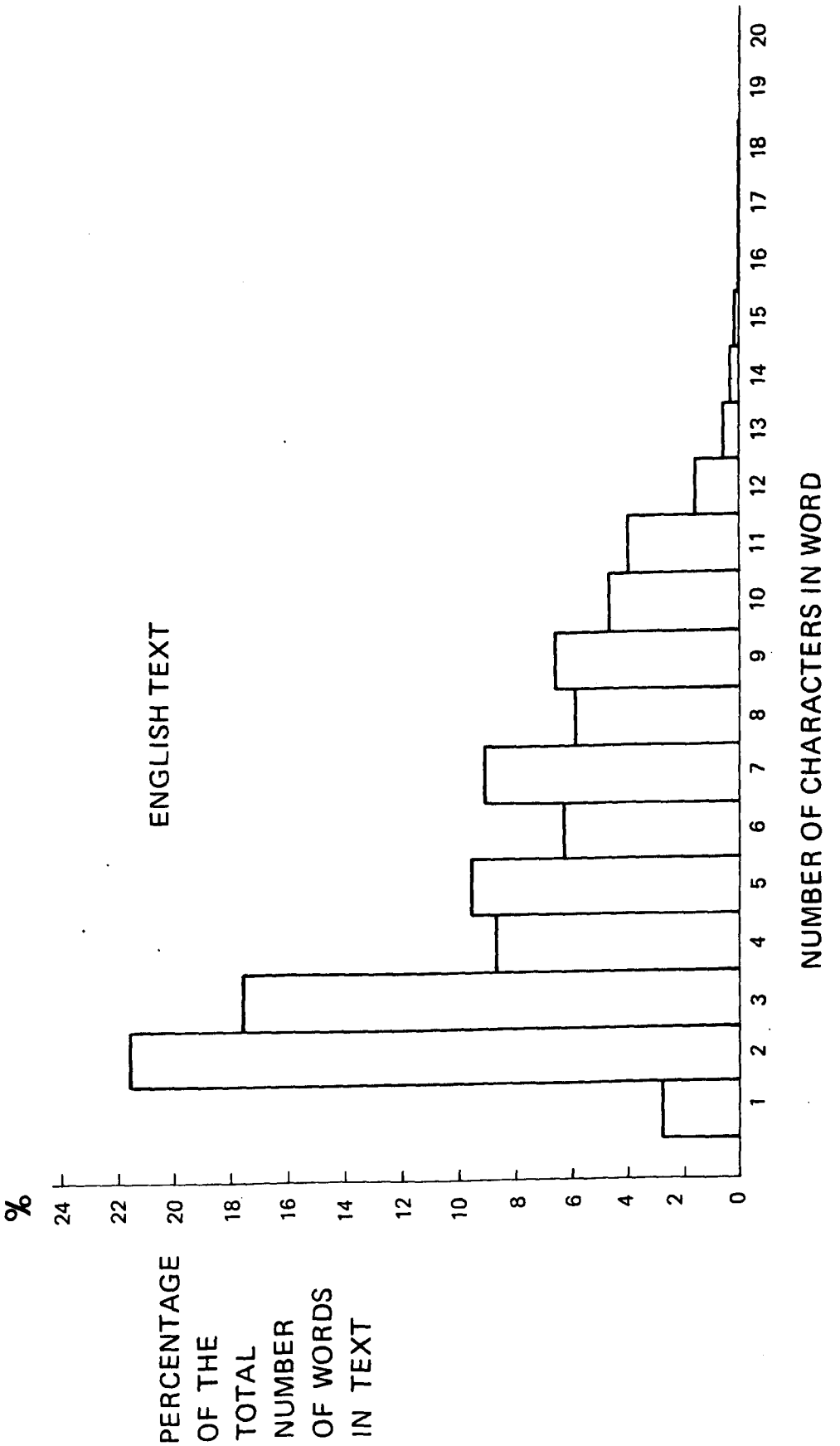


FIGURE VI.2(b)

HISTOGRAM SHOWING THE NUMBER OF CHARACTERS IN A WORD AS A FUNCTION OF THE TOTAL NUMBER OF WORDS

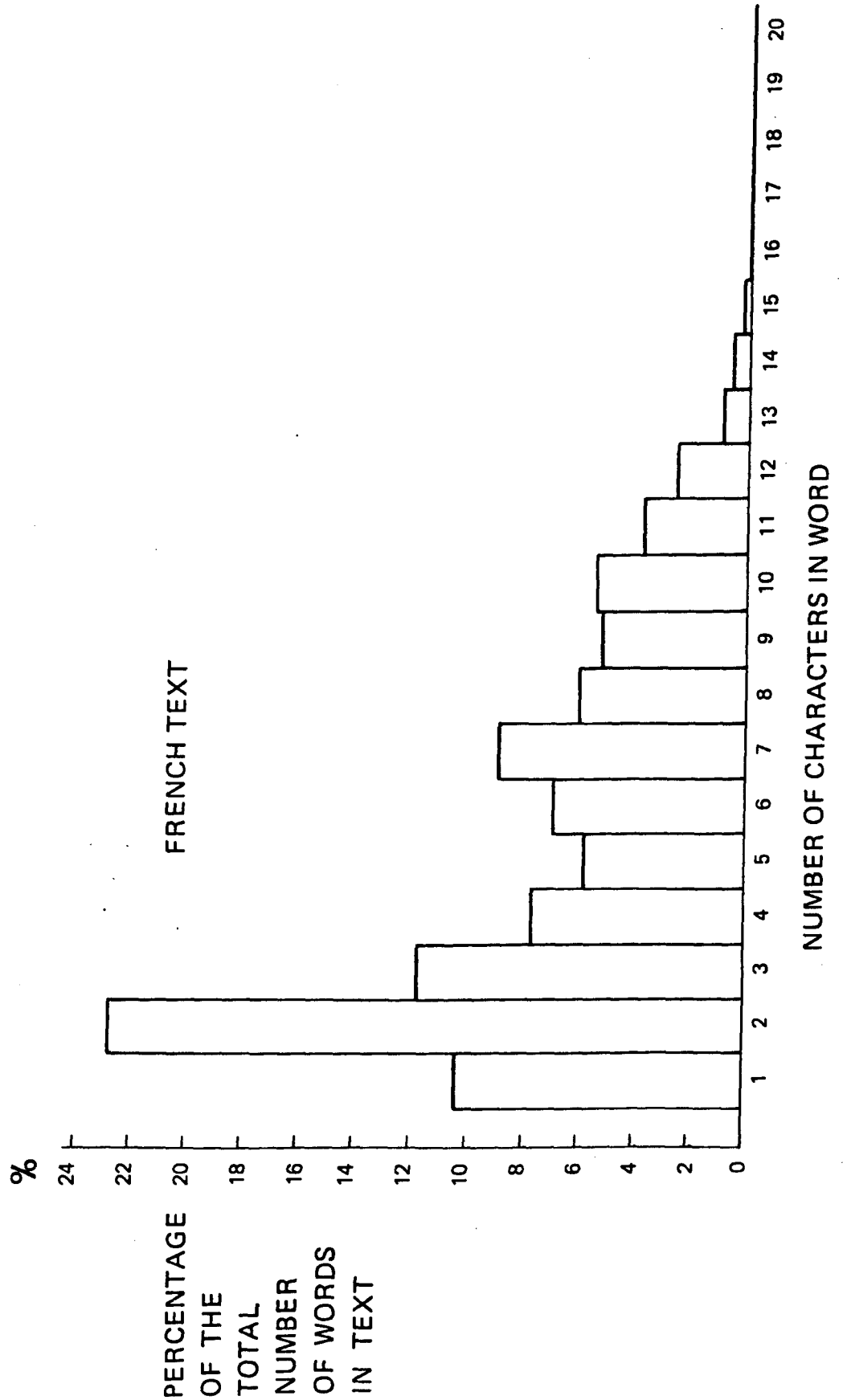


FIGURE VI.2(c)

of one character are far more prevalent in French, there being 10.4% compared with 2.8% in English. Apart from the section labels used widely in this particular data base, the only single letter words in English are 'A', 'I' and 'O', and only the first two occur in the Treaties. However, several single letters constitute words in French, according to the definition of a word given in section IV.2.2. In particular, any character string followed by an apostrophe is a word, so, for example, 'L' and 'D' as used in L'Europe and D'une constitute words. These particular two examples are common words and account for approximately 5.9% of the total number of words.

The most frequently occurring length in both languages is two characters, which is to be expected since many of the common words are two letter words. Three letter words are far more frequent in English, 17.6% of all the English words having three letters compared with 11.8% of the French. This high percentage is due mainly to the word 'The', which is the most frequent word in English and accounts for about 10% of the text. On the other hand the most frequent three letter word in the French text, the word 'Des', occurs only 5920 times compared with 26,322 occurrences of 'The', and constitutes only 2% of the French text.

HISTOGRAM SHOWING THE NUMBER OF CHARACTERS
IN A WORD AS A FUNCTION OF THE NUMBER OF DIFFERENT WORDS

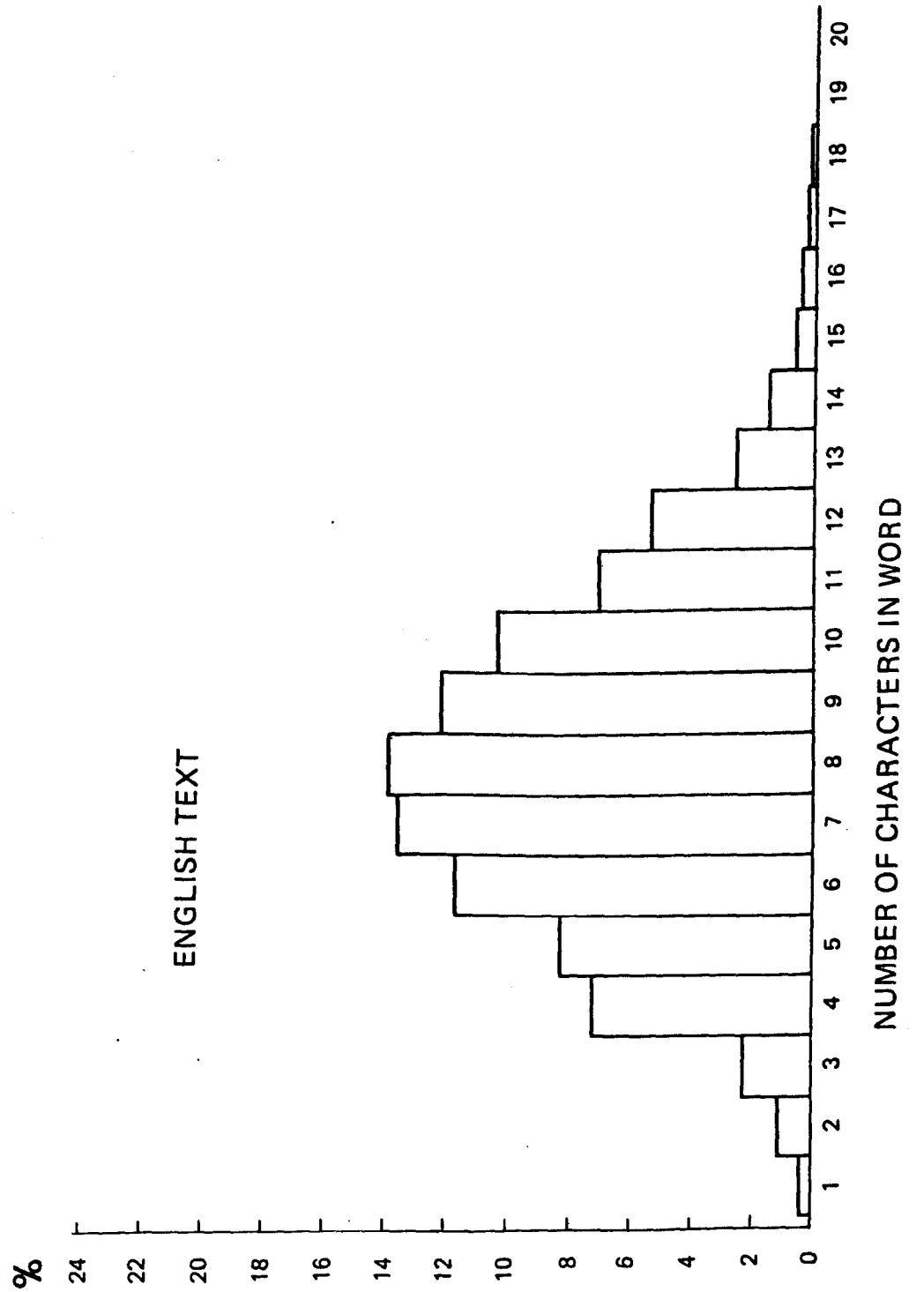


FIGURE VI.2(d)

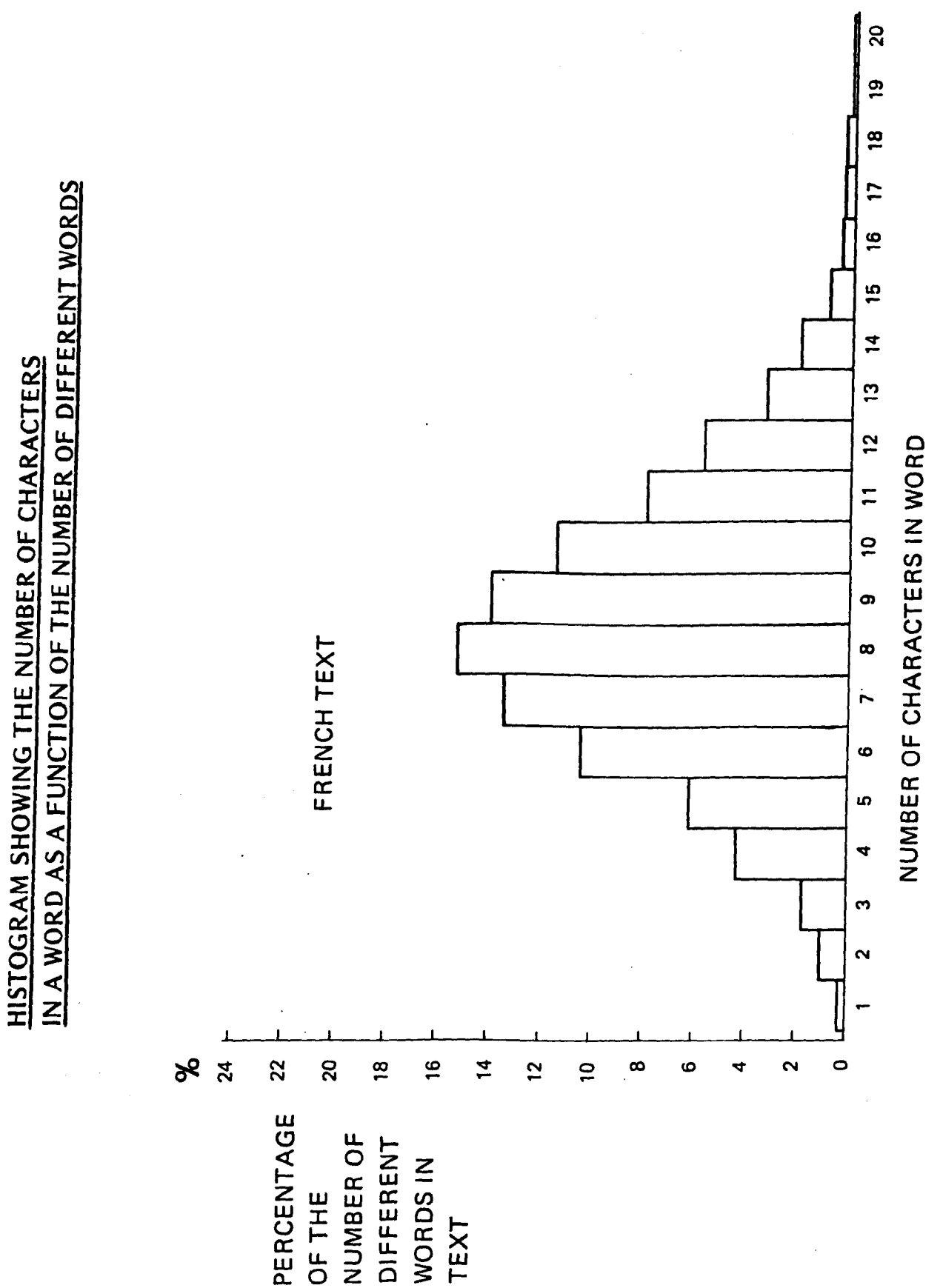

HISTOGRAM SHOWING THE NUMBER OF CHARACTERS
IN A WORD AS A FUNCTION OF THE NUMBER OF DIFFERENT WORDS
FRENCH TEXT
PERCENTAGE OF THE NUMBER OF DIFFERENT WORDS IN TEXT
%
0
2
4
6
8
10
12
14
16
18
20
22
24
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
NUMBER OF CHARACTERS IN WORD

FIGURE VI.2(e)

The remaining lengths are distributed fairly similarly, except that words of five and nine characters are both noticeably more prevalent in English.

As is obvious from the above discussion, the distribution of word lengths as a percentage of the total number of words is determined to a certain extent by the lengths of the most frequently occurring words. To analyse word length as a feature of the vocabulary, in contrast to the usage of words, we examine the percentage of different words having a particular length. These distributions are shown in Figures VI.2(d) and VI.2(e).

The distribution of word lengths is different from the previous analysis, the shorter words being much less prominent. The length most common in both languages is eight letters. In general, words longer than eight letters are more prevalent in French, whereas English tends to favour shorter words of four, five and six letters.

VI.2.4 Zipf's Law for single words

The graphs labelled A in Figures VI.2(f) and VI.2(g) verify that the two texts satisfy the relation known as Zipf's Law, see Zipf (58), which says that the frequency of occurrence of a word is proportional to some negative power of its rank, when the words are ranked in decreasing order of frequency. However, the relation holds for the middle

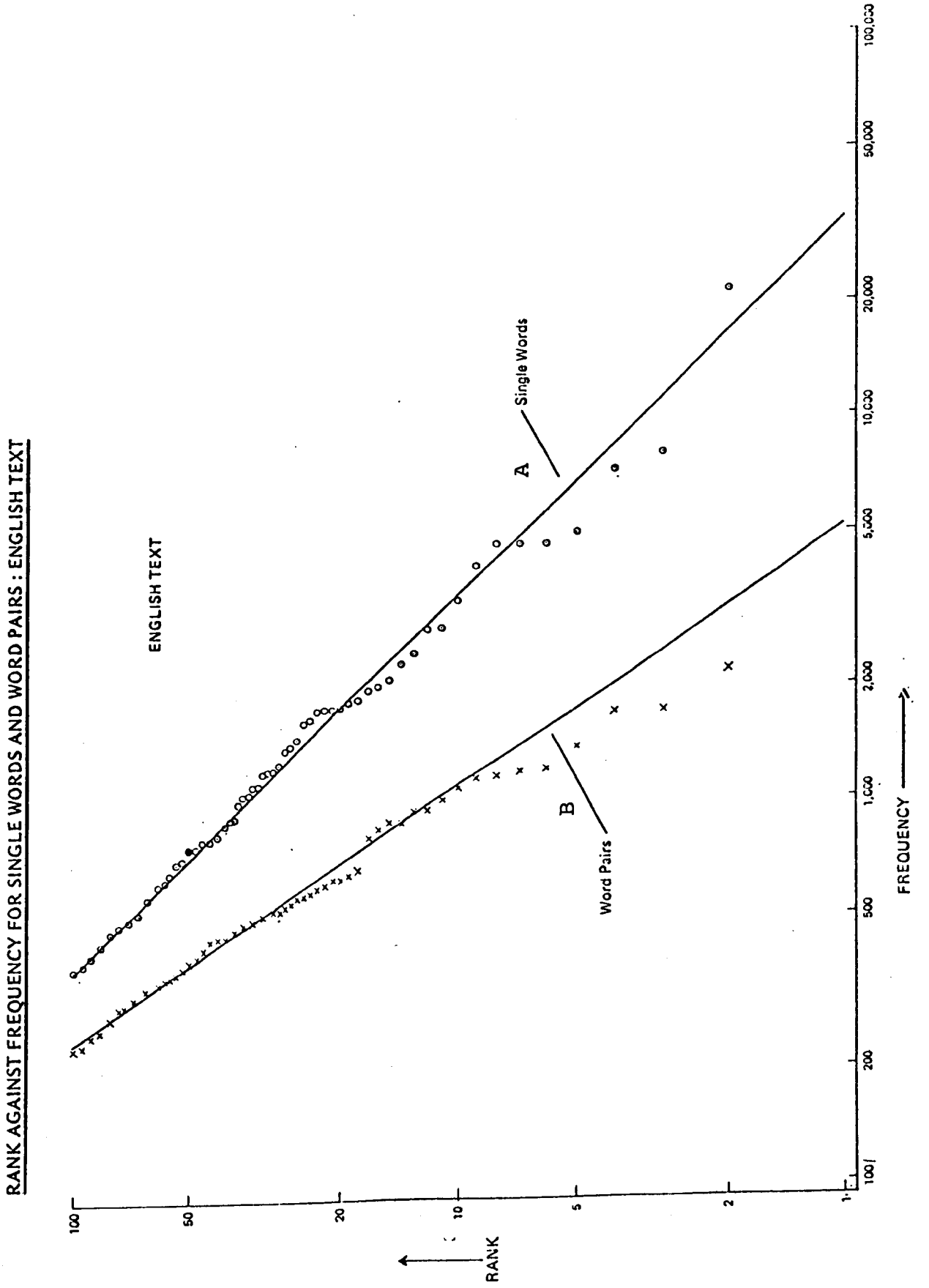

RANK AGAINST FREQUENCY FOR SINGLE WORDS AND WORD PAIRS : ENGLISH TEXT
ENGLISH TEXT
A
Single Words
B
Word Pairs
RANK
FREQUENCY
100
50
20
10
5
2
1
100
200
500
1,000
2,000
5,000
10,000
20,000
50,000
100,000

FIGURE VI.2(f)

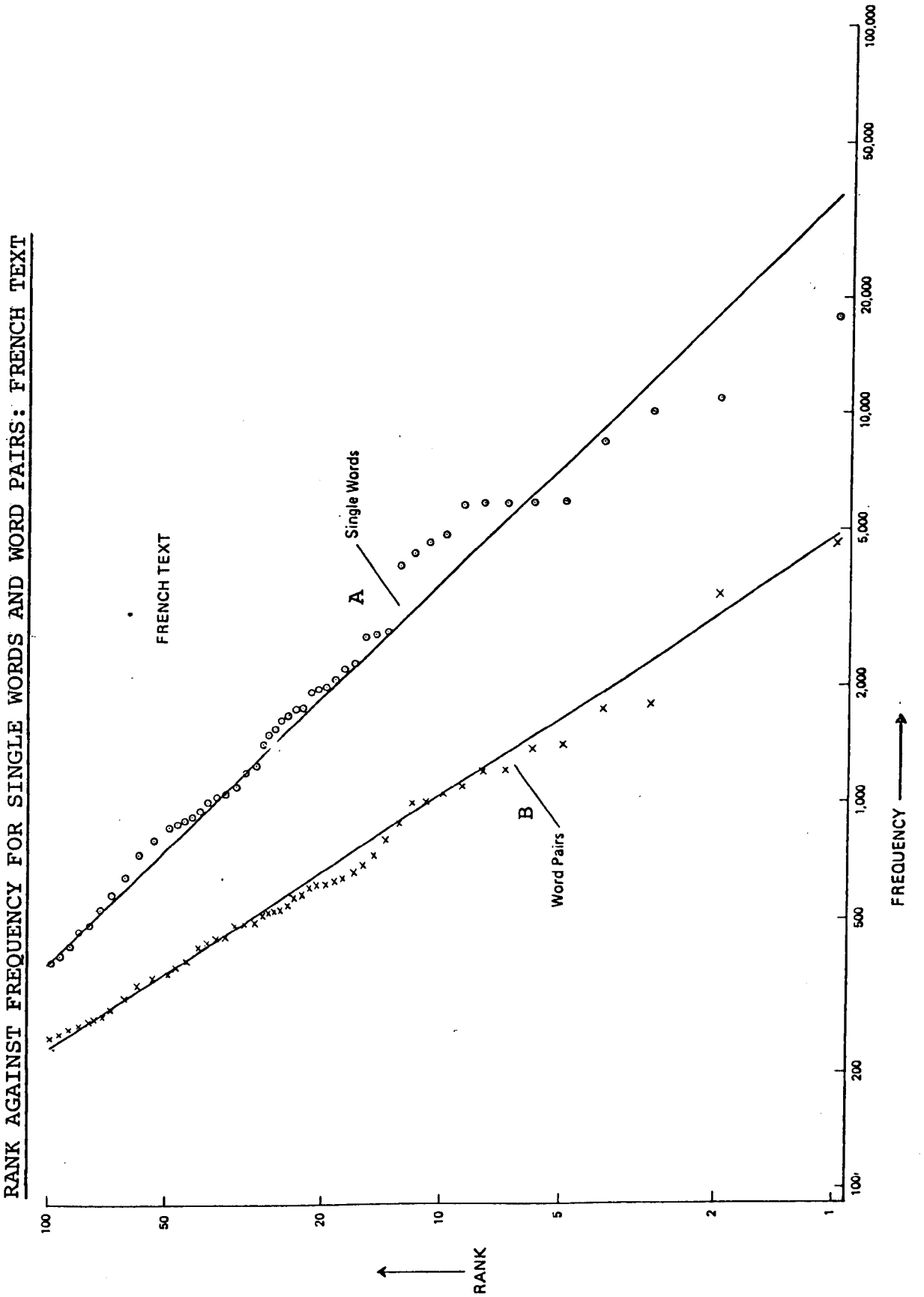

RANK AGAINST FREQUENCY FOR SINGLE WORDS AND WORD PAIRS: FRENCH TEXT
FRENCH TEXT
A
Single Words
B
Word Pairs
RANK
FREQUENCY
100
50
20
10
5
2
1
100
200
500
1,000
2,000
5,000
10,000
20,000
50,000
100,000

FIGURE VI.2(g)

range of frequencies only, and not for the very frequent and very rare words. For this middle range the graphs can be approximated to straight lines given by the equation

$$\log r = A \log f + B$$

where f = frequency of occurrence of word

r = rank of word

and A and B are constants.

The relation can also be represented in the form

$$f = K r^{-a}$$

where K and a are constants, and in this case a is approximately equal to 1.

VI.3 Statistics of word pairs

Table VI.3(a) gives the total number of pairs of adjacent words in the two texts, and the number of these pairs which are different. These values have been determined for the two cases of including and excluding the common words.

Table VI.3(a) Number of word pairs used in Treaties

		English text	French text
Common words included	Total number of word pairs	274,451	287,599
	Number of different word pairs	54,856	60,726
Common words excluded	Total number of word pairs	136,037	137,981
	Number of different word pairs	61,356	64,601

At first it seemed surprising that there were more different pairs when common words were excluded, as the reverse is obviously true for the total number of pairs. However, when the common words are included many of the pairs consist of one non-common word preceded or followed by a common word. This restricts the number of different pairs, as there are a limited number of these common words. Removing the common words results in a greater variety of adjacent words. To illustrate this behaviour we consider the following four sentences:

Austria and Belgium and Cyprus and Denmark have accepted.

Belgium and Austria and Denmark have ratified.

Belgium and Denmark and Cyprus have accepted.

Denmark and Austria and Cyprus have ratified.

Taking all words into account we get the following 12 pairs:

Austria - and	and - Cyprus
Belgium - and	and - Denmark
Cyprus - and	Cyprus - have
Denmark - and	Denmark - have
and - Austria	have - accepted
and - Belgium	have - ratified

If on the other hand we regard 'and' and 'have' as common words and exclude them, we get the following pairs:

Austria - Belgium	Denmark - Austria
Austria - Cyprus	Denmark - Cyprus
Austria - Denmark	Cyprus - accepted
Belgium - Austria	Cyprus - ratified
Belgium - Cyprus	Denmark - accepted
Belgium - Denmark	Denmark - ratified
Cyprus - Denmark	

In this case there are a total of 13 different word pairs.

VI.3.2 The most frequent word pairs

Tables VI.3(b) and VI.3(c) list the 20 most frequent word pairs in the English and French text respectively. The pairs have in this case been formed from all words including the common words.

Table VI.3(b) The 20 most frequent English word pairs

OF	- THE	7082	FOR	- THE	997
TO	- THE	2175	PROVISIONS	- OF	943
SHALL	- BE	1720	THE	- PROVISIONS	935
IN	- THE	1709	OF	- A	869
THE	- COUNCIL	1374	OF	- ARTICLE	866
COUNCIL	- OF	1214	THIS	- CONVENTION	839
OF	- EUROPE	1196	WITH	- THE	801
BY	- THE	1156	IN	- RESPECT	665
CONTRACTING	- PARTY	1137	RESPECT	- OF	639
OF	- THIS	1075	ACCORDANCE	- WITH	623

Table VI.3(c) The 20 most frequent French word pairs

DE	- L	4597	PRESENTE	- CONVENTION	993
DE	- LA	3434	PARTIE	- CONTRACTANTE	989
A	- L	1787	L	- ETAT	872
L	- ARTICLE	1737	SECRETAIRE	- GENERAL	794
A	- LA	1407	D	- UN	726
DU	- CONSEIL	1366	LA	- CONVENTION	687
L	- EUROPE	1211	OU	- D	657
CONSEIL	- DE	1199	PARTIES	- CONTRACTANTES	633
D	- UNE	1091	GENERAL	- DU	612
LA	- PRESENTE	1043	DE	- RATIFICATION	611

With the exception of the French pair 'Secretaire General', all these most frequent pairs contain at least one common word. The equivalent pair 'Secretary General' is not among the 20 most frequent English pairs. This is because of an inconsistency whereby the pair is sometimes hyphenated to produce a single word, and thus the frequency of the pair is reduced to a count of the unhyphenated occurrences.

One very frequently used phrase is 'Council of Europe'. From Table VI.3(b) we see that the pair 'Council of' occurs 18 times more than the pair 'of Europe'. Thus there are occurrences of 'Council of' which are not immediately

followed by the word 'Europe'. An example occurs in the sentence:

> 'The Secretary General of the Council of Europe shall notify the members of the Council of the date of entry into force of this Convention.'

In this context the underlined pair is not meaningful, as the two words are not syntactically related. However, all occurrences of the pair 'Council of' are treated as equivalent in document vectors.

VI.3.3 Zipf's Law for word pairs

The graphs labelled B in Figures VI.2(f) and VI.2(g) demonstrate the relation between the rank and frequency of word pairs. Again the middle range of values obey Zipf's Law, though the shape of the graph is different in this case. In the equation $f = K\ r^{-a}$ representing the straight line approximation the value of a is approximately 2/3.

CHAPTER VII

CLUSTER ANALYSIS IN COURT

VII.1 Introduction

During the course of this study an opportunity arose to investigate a novel application of cluster analysis, namely to test the trustworthiness of certain oral confessions presented as evidence in a criminal trial.

The defendant was alleged to have made certain statements to the police, which he denied having made. It occurred to him that a computer analysis of the statements might prove that the words were not his own, by revealing differences in style between the statements and his own speech. The style of speech or text is determined to a certain extent by the words used, and the analysis of the word content of documents is precisely the concern of this study. Hence it seemed appropriate to use the technique of clustering developed here to compare the style of the disputed statements with the defendant's own style. (Clustering has been used successfully in the past by Ule (59) and (60) for studying the style of literary texts, in particular for comparing Elizabethan plays.)

In addition to the disputed statements, the defendant was asked to supply some text which he could prove was his own, so that a comparison could be made. A set of statements

made by him in a previous trial were available, and these had been made under circumstances identical to those under which the disputed statements were alleged to have been made. If the two sets of statements differed in style we would expect them to cluster into separate groups, and this hypothesis was tested as described in the following sections.

VII.2 Discriminators of style

The so-called function words, articles, prepositions and conjunctions for example, are generally agreed to characterise the style of a speaker or writer, and therefore indicate authorship. The use of such words is independent of the subject matter of the text, and is therefore optional to a certain extent. In the Hebrew language, for example, the definite article is almost always optional, and Radday and Shore (61) have found it to be a 'valid and quite powerful discriminant' in their study of the books of the Hebrew Bible. Morton and Winspear (62) have studied the occurrence of the word 'kai' as an indicator of the authorship of Greek texts.

In English there is no one particular word which is sufficiently optional to be a good discriminator on its own, and it is more usual to consider the use of a large set of function words. Of course the grammatical structure of a sentence determines, to a certain extent, the function words

to be used, but often there is a choice of structures for expressing the same idea, and this choice lies with the author. For example, the two phrases

'Analysing style by computer'

and

'The analysis of style by computer'

convey the same meaning, though the second contains the two function words 'the' and 'of' which the first does not. An author who consistently prefers the second construction will use these two words more frequently.

Mosteller and Wallace (63) in their investigation of the disputed authorship of certain Federalist papers used function words, which they call 'filler' words, as discriminators of style. In reference (60), Ule's discriminators consisted of the four most common words in the texts being studied, 'and', 'the', 'to' and 'I'.

VII.3 Classification of the statements

A selection was made of function words to be used as discriminators of style, and these are listed in Table VII.3 in alphabetical order. These words were used for the vector representations of the statements. That is, the vector elements were the frequencies of occurrence of the function words, the content words being excluded. These vectors were clustered in the usual way.

Table VII.3 Function words used as discriminators of style

A	FROM	MUST	THEY
ABOUT	GET	MY	THIS
AM	GOT	NOT	TO
AN	HAD	OF	UNTIL
AND	HAVE	OFF	UP
ANY	HE	ON	US
ARE	HER	OR	WAS
AS	HIM	SHE	WE
AT	HIS	SHOULD	WERE
BE	I	SINCE	WHAT
BECAUSE	IF	SO	WHEN
BEEN	IN	SOME	WHICH
BEFORE	INTO	THAT	WILL
BUT	IS	THE	WITH
BY	IT	THEM	WOULD
CAN	ITS	THEN	YOU
DO	MAY	THERE	
FOR	ME	THESE	

The dendrogram representation of the clustering is shown in Figure VII.3. Documents 1 - 7, the undisputed documents, contain the statements which the defendant agreed he had made in his own words. The disputed documents, 8 - 11, are the statements which the defendant was alleged to have made but which he denied.

We see from Figure VII.3 that the disputed documents have clustered together in one group, separately from the undisputed ones. It is not obvious, however, from the diagram that documents 1 - 4 are more similar to 5, 6 and 7 than to 8, 9, 10 and 11. We must examine the similarity coefficients for further confirmation of the dissimilarity of the two sets of statements. These coefficients reveal that where each of 1 and 2, and the cluster of 3 and 4 have joined the merged cluster of documents 5 - 11, they have done so by way of their similarity to one of 5, 6 or 7, the other undisputed documents.

The way in which the undisputed documents have clustered shows that the defendant uses a wide variety of the function words chosen, since some of the coefficients of similarity between the documents in this set are quite low. The disputed documents, in contrast, form a tight cluster at a relatively high level of similarity, which suggests a different overall use of the function words.

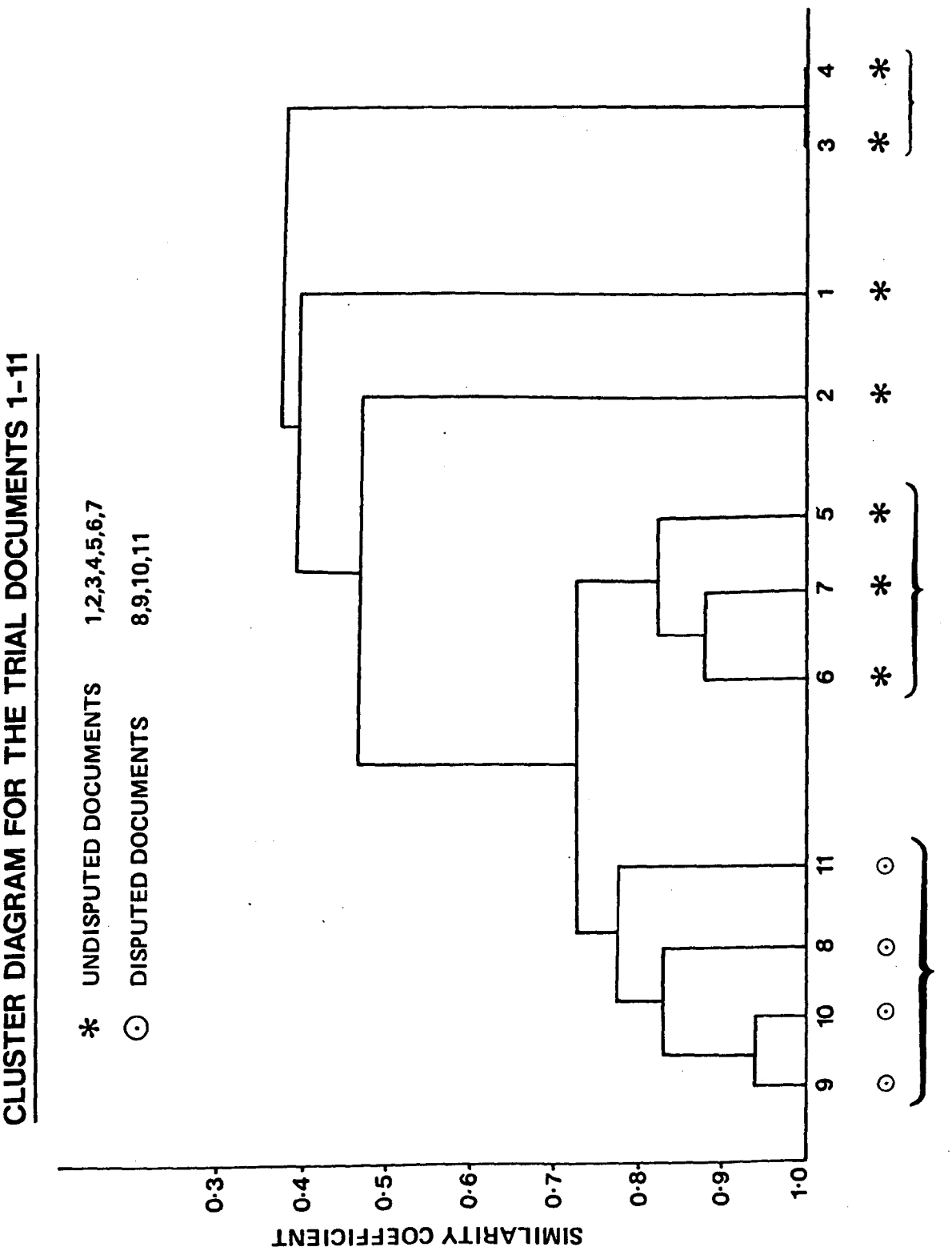

CLUSTER DIAGRAM FOR THE TRIAL DOCUMENTS 1-11
* UNDISPUTED DOCUMENTS 1,2,3,4,5,6,7
⊙ DISPUTED DOCUMENTS 8,9,10,11
SIMILARITY COEFFICIENT
0·3
0·4
0·5
0·6
0·7
0·8
0·9
1·0
4
3
1
2
5
7
6
11
8
10
9

FIGURE VII.3

VII.4 Comparative tests

To test the reliability of the method of cluster analysis based on a set of function words as discriminators of style, the technique was applied to some similarly-sized collections of documents whose authorship was known.

VII.4.1 Analysis of sonnets

The first test collection consisted of twelve sonnets, six by Shakespeare and six by Wordsworth. These were converted to vector form using the function words in Table VII.3 and excluding all other words. The vectors were clustered and Figure VII.4(a) is the resulting dendrogram.

Documents 1 - 6 are the Shakespeare sonnets and documents 7 - 12 those by Wordsworth. It is apparent that no well-defined clusters discriminate between the two sets of documents. Thus the function words cannot distinguish the two poets in relation to their sonnets. This is not surprising, because sonnets have a well-defined structure which partly determines the function words to be used. The similarity in style is deliberate on the part of the poets.

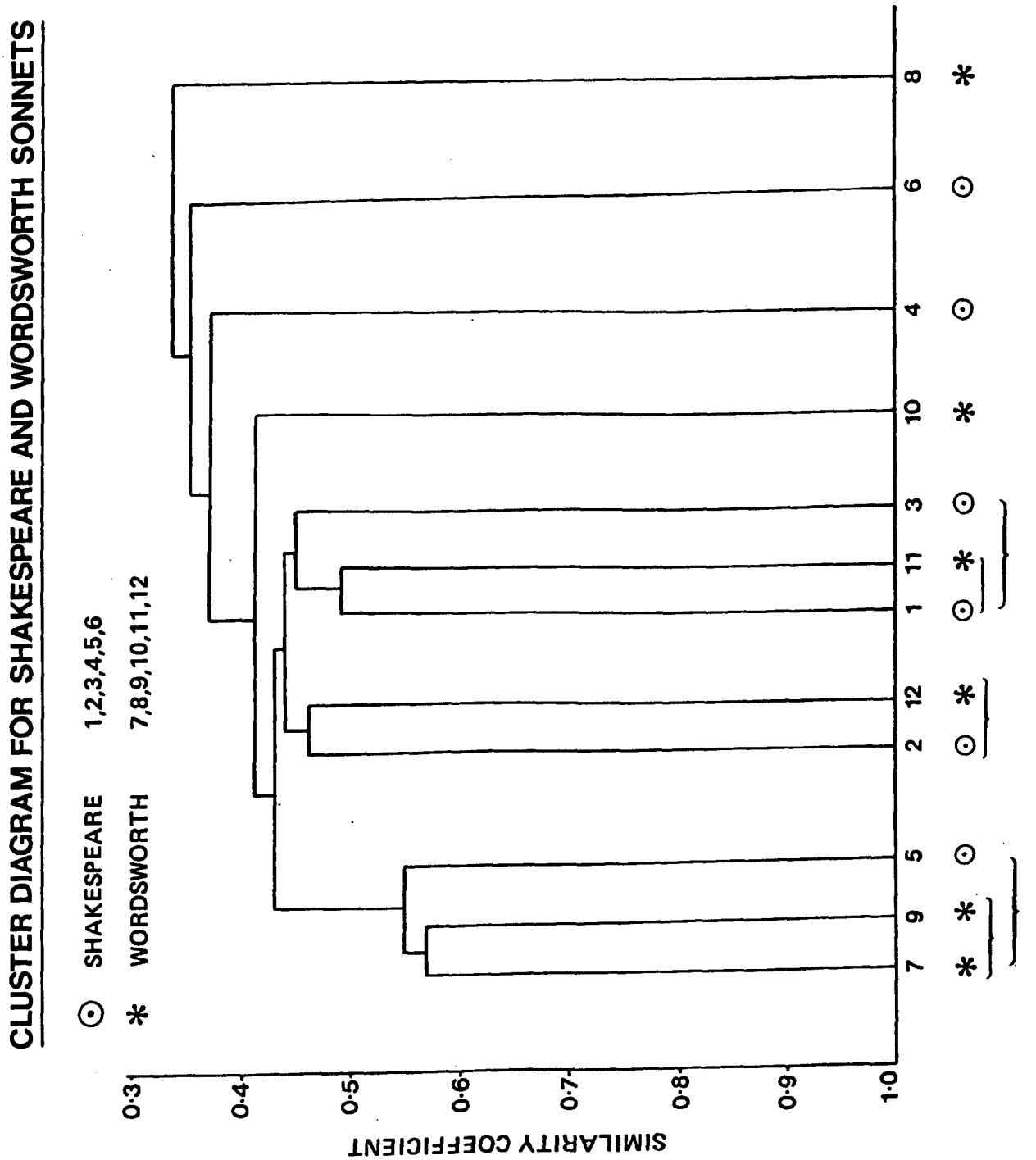


FIGURE VII.4(a)

VII.4.2 Analysis of collects and extracts from a novel

A further set of twelve documents was used for a second test of the method. These documents consisted of six collects from the Book of Common Prayer, and six extracts from the novel 'Guys and Dolls' by Damon Runyon. The same procedure was adopted as above, and the dendrogram is shown in Figure VII.4(b).

In this test the method discriminates between the two sets of documents extremely well. Two distinct clusters are formed, separating the two different styles.

VII.5 The verdict

From the comparative tests we can conclude that the function words chosen are good discriminators of style, when clustering is based on their frequency of occurrence. They have differentiated the archaic English of the collects from the comparatively modern American gangster-like language of the Runyon novel. On the other hand they have rightly failed to detect any difference in the style of a collection of sonnets. Thus we can say that the separation of the disputed and undisputed statements according to their use of function words indicates a difference in their style. This difference is not as great as that between the collects and the novel, as the statements all employ language in use at the same period, that is, within the same decade.

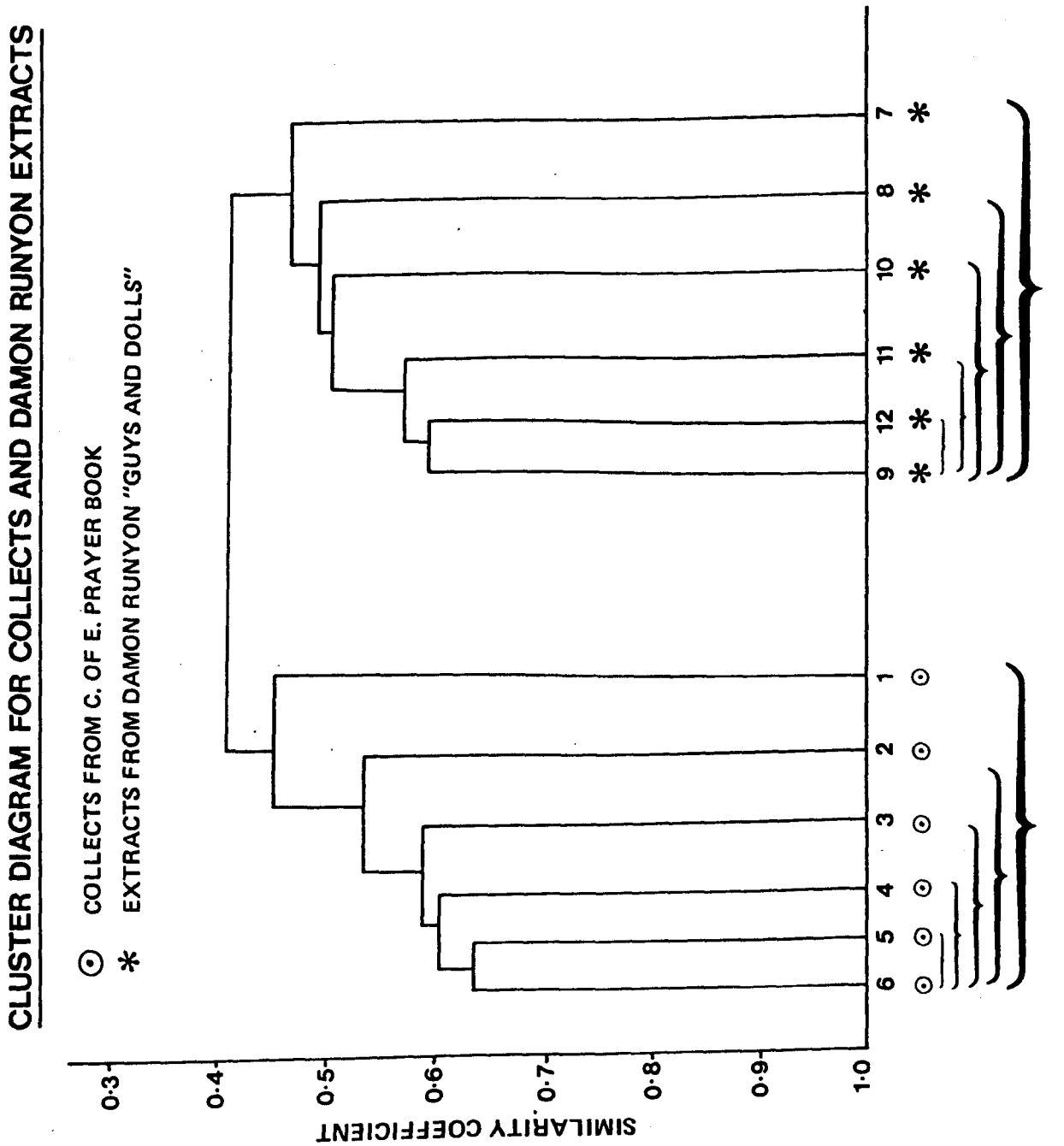


FIGURE VII.4(b)

Figures VII.3, VII.4(a) and VII.4(b) were exhibited in the criminal proceedings, and the expert evidence given to the court was that on the basis of the above analysis, the disputed statements were unlikely to have been made by the defendant. He was eventually acquitted on those counts in the indictment to which the alleged statements were of relevance.

VII.6 Conclusion

This is believed to be the first occasion that cluster analysis has been used in a criminal trial to provide evidence on the authenticity of statements alleged to have been made by a defendant. However, it is unlikely that this type of expert evidence will be used often. In this particular case the circumstances were ideal; two sets of documents of adequate length were available, and which were produced, or were alleged to have been produced, under identical conditions. The occasions on which such documents will be available for cluster analysis are likely to be rare. However, this report is included here to demonstrate a further practical application of the technique of cluster analysis as developed in this study.

CHAPTER VIII

CLASSIFICATION OF STATUTES

BASED ON CITATIONS

VIII.1 Introduction

This chapter discusses a further application of the technique of cluster analysis which exploits an important feature of legal documents, namely the marginal citations found in Statutes. These citations are shorthand references to U.K. Statutes cited in full in the main body of the text of a given Statute, and which are directly relevant to that Statute.

Citations in legal documents are analagous to the bibliographic references quoted in scientific papers which have been used successfully for the classification and retrieval of such documents. For example, Price and Schiminovich (64) have used the co-occurrences of bibliographic citations to measure the relatedness of articles on theoretical high energy physics in order to establish a classification scheme. Michelson et al (65) have investigated the use of bibliographic data in a vector search system to provide additional search terms. They found that searching on citations alone was as effective as searching on subject keywords, and suggested that the two types of document

descriptors could be used in combination to improve retrieval performance. Salton's experiments (66) provide additional verification of the usefulness of bibliographic information for document retrieval.

Following the success of bibliographic references as document descriptors in scientific subjects, we would expect this technique to work as well, if not better, in the field of Law, where the use of citations is more highly established than in any other discipline. In scientific documents, references are included at the discretion of the author, according to his awareness of relevant papers, and to the age of this literature. In his study of scientific citation networks, de Solla Price (67) discovered that the citation rate drops markedly for documents more than 10 years old, and that very new documents are unlikely to be cited at all, since they have not had time to be noticed. In contrast, legal documents may still be relevant after 100 years or more and will be cited where necessary, and, further, legal draftsmen must be aware of the very latest legislation.

The following sections describe the classification of a collection of Statutes based on the occurrence of marginal citations.

VIII.2 Citation data

VIII.2.1 Extraction of citations

The document collection used for this experiment consisted of the Public General Acts of the United Kingdom Parliament for the three years 1973 - 1975. The marginal citations were extracted from the printed versions of these documents, as the text is not yet available in machine readable form, although H.M.S.O. expects to produce magnetic tape versions by 1980.

At first the task of extracting the citations seemed simple; marginal notes are clear and printed in smaller type than the main body of the text, and should have been readily identifiable and easily separable. However, on a closer inspection of the marginal notes, several inconsistencies and errors were found. Some citations were incomplete as to either the year or the chapter number. These were relatively easy to identify and complete, by extracting the information from the corresponding reference in the text, though in the case of a missing chapter number this might be available from outside sources only, if the same Act were not cited correctly elsewhere in the text.

Some references within the text had no marginal citation opposite, whilst some others which did were given an additional citation where the text of the relevant section ran onto a separate page. Missing citations were inserted, and additional citations for separate pages were deleted.

Inconsistencies also occurred in the number of marginal citations given for a repeated reference. Sometimes one citation only was given to cover all occurrences of the same reference in a given section. In other cases the citation was repeated for each individual reference. For this experiment just the citations actually quoted in the marginal notes were used, in the case of repeated references. However, future work is expected to examine more closely the principles for selecting citations, and for weighting purposes it might be more effective to include additional citations for multiple occurrences of a reference.

The extracted citations, completed and corrected where necessary, were used to compile a new data base in which each original Act was replaced by the appropriate set of citations. In addition, each document was assigned a reference to itself, since it was considered desirable to be able to associate a given document with any other document which cited it, by the occurrence of the citation for the given document.

VIII.2.2 Representation of citations

Each citation of an Act was represented by seven numeric digits preceded or followed, in some special cases, by certain alphabetic characters. The first four digits consisted of the year number, and the remaining three were given by the chapter number of the Act, e.g. 1975037.

Pre-union Scottish Acts were indicated by an 'S' after the chapter number, e.g. 1597240S. The letters 'NI' following the chapter number, as in 1972009NI, were used to indicate an Act or Statutory Instrument of the Northern Ireland Parliament.

The year numbers of Local Acts, Statutory Rules and Orders, Orders and Statutory Instruments were preceded by the letters 'L', 'SR', 'O' and 'SI' respectively, and followed by a four figure number. In the case of a Local Act this number represented the chapter number, converted from the small Roman numeral original, and in the case of the latter three, the registration number.

Where EEC Council Regulations were referred to in an Act, these were cited in the margin by the reference to the Official Journal number with an alphabetic reference and a page number. These citations were represented by the letters 'OJ' followed by the alphabetic reference plus four digits for the page number.

VIII.3 Classification of Statutes

There were 210 Statutes in the document collection, and these contained a total of 4390 citations, including the additional citations by which each document referred to itself, and 1110 of these were distinct. The document collection in

citation form was used as input to the suite of classification programs (see Chapter IV), each citation being regarded as a word of text. There were no 'common words' in this case.

Of the 21,945 similarity coefficients calculated, 3,315 , that is approximately 15%, were non-zero. The highest coefficient was 0.891, the value of similarity between the Pensioners' Payments Acts for 1974 and 1975.

The most significant clusters were formed at or above the level of similarity of 0.247. At this level, 65 of the 210 documents had been assigned to clusters. The dendrogram showing the clustering as far as level 0.247 is given in Figure VIII.3. The meaningful, well-formed clusters have been labelled in the diagram, and the headings which correspond to these labels are listed in Table VIII.3, together with the number of documents in each cluster.

The document numbers shown in the diagram are formed from the year and chapter numbers for each Act. For example, 7405 is Act number 5 enacted in 1974. (Note that we can use a shorter notation than that required for the citations themselves, since the documents are all 20th Century Acts, so that no ambiguity arises from omitting '19' from the year numbers. Two digits suffice for the chapter numbers since the maximum number of Acts enacted in any of the three years studied is 83.) A list of the titles of the Acts which correspond to these document numbers is given in Appendix D.

Table VIII.3 Headings for clusters of Statutes

A	Statutory corporations (Financial provisions) (2)
B	Independent broadcasting authority (2)
C	Ministerial matters (2)
D	Disqualification of Members of Parliament (2)
E	Representation of the people (2)
F	Family matters (2)
G	Northern Ireland (2)
H	Health (2)
I	Criminal matters (2)
J	Pensioners' payments (2)
K	Social security (i) (4)
L	Social security (ii) (3)
M	Housing (Scotland) (2)
N	Housing (3)
O	Industry (2)
P	Local government (2)
Q	Town and country planning (4)
R	Scottish development (2)
S	Land compensation (2)
T	Insurance companies (2)
U	Northern Ireland (Emergency provisions) (3)
V	Finance (4)
W	Dogs (2)
X	British and Commonwealth nationality (2)
Y	Appropriation (4)
Z	Export guarantees (2)

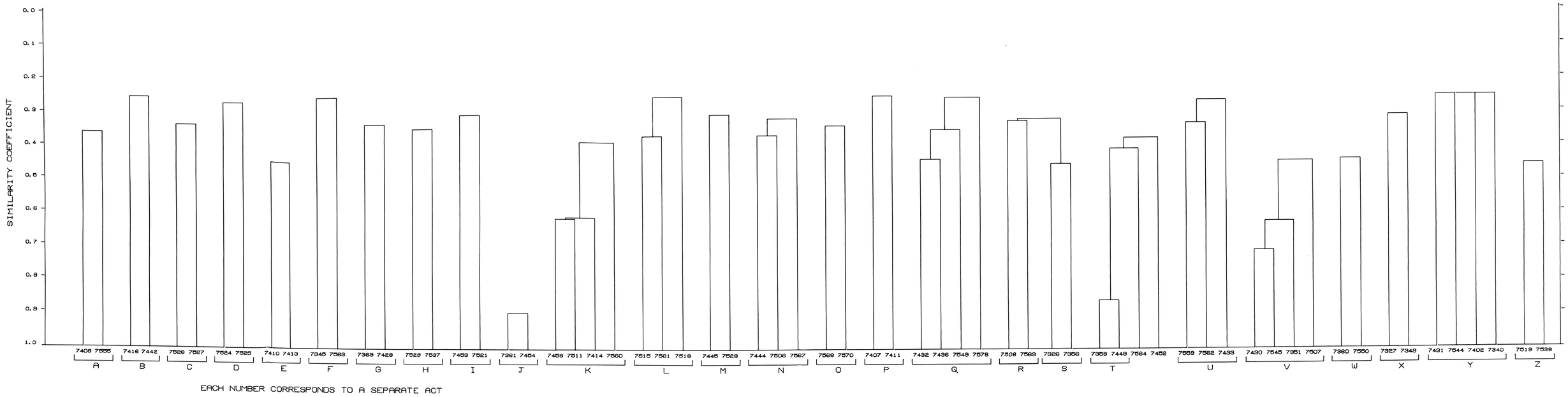
DENDROGRAM SHOWING

CLASSIFICATION OF

ACTS

BASED ON CITATIONS.

FIGURE VIII.3


SIMILARITY COEFFICIENT
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
7408 7555
7416 7442
7526 7527
7524 7525
7410 7413
7345 7563
7369 7428
7529 7537
7453 7521
7361 7454
7458 7511 7414 7560
7515 7561 7518
7445 7528
7444 7508 7567
7568 7570
7407 7411
7432 7436 7549 7578
7508 7569
7326 7356
7358 7449 7564 7452
7559 7562 7433
7430 7545 7351 7507
7360 7550
7327 7349
7431 7544 7402 7340
7519 7538
A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
P
Q
R
S
T
U
V
W
X
Y
Z
EACH NUMBER CORRESPONDS TO A SEPARATE ACT

Four of the best-formed clusters are K, Q, V and Y on 'Social Security', 'Town and Country Planning', 'Finance' and 'Appropriation' respectively, and we examine these in more detail.

Cluster K is joined by the National Insurance and Supplementary Benefit Act for 1973 at level 0.191 not shown in the dendrogram. Later, at level 0.131, that cluster merges with the other cluster of Social Security Acts shown in the dendrogram, cluster L.

Cluster Q consists of the Acts on Airports Authority, Mobile Homes, Mines Working Facilities and Support, and Town and Country Amenities, which all deal with various aspects of town and country planning. They have been brought together by the mutual citation of the Town and Country Planning Act 1971, and the Town and Country Planning (Scotland) Act 1972. These two Acts are also responsible for most of the chaining occurring below level 0.247, which consists of Acts chaining onto the 'Town and Country Planning' cluster.

As expected the four Finance Acts in the document collection (there are two for 1975) have clustered together. This cluster remains unaltered until level 0.072, when it is joined by the cluster consisting of the Counter Inflation Act 1973, the Rate Rebate Act 1973 and the General Rate Act 1975, which deal with financial matters.

The 'Appropriation' cluster is formed by the four Appropriation Acts (two for 1974) all joining at level 0.25. Each of these Acts contains just one citation to another Act which is the Public Accounts and Charges Act of 1891 in each case. In addition each document contains a reference to itself which is of course unique for each one. The matching pattern for each pair of Appropriation Acts is therefore the same; one citation matches and one does not. Hence the coefficient of similarity is the same for each pair. On examining the clustering below level 0.247, the 'Appropriation' cluster is found to be disjoint from all other clusters except at level zero where everything merges. The Appropriation Acts deal with the means for appropriating funds for government use, and are not normally cited by other Acts.

Certain spurious clusters are formed below level 0.247. For example, the Farriers (Registration) Act 1975 and the Moneylenders (Crown Agents) Act 1975 are joined at level 0.167. These Acts are not semantically related, but have been brought together by the Interpretation Act of 1889, which interprets the meaning of certain terms used in these Acts. In future experiments, this and other general Acts might usefully be designated 'common' and excluded, as they have no direct bearing on the semantic content of the Acts which cite them.

VIII.4 Conclusion

The Acts of Parliament for 1973 - 1975 have been clustered successfully on the basis of the marginal citations they contain. Semantically related Acts have been grouped together, and above a level of similarity of 0.247 most of the clusters are highly homogeneous.

The data for the clustering was extracted manually from the printed Acts, though it is hoped that in the future it will be possible to extract the citations automatically from magnetic tape versions of the Acts. The main difficulties will lie in overcoming the problems of erroneous citations discussed in section VIII.2.1.

Further investigation is needed into the use of marginal citations as an aid for searching legal documents by computer.

CHAPTER IX

VECTOR SEARCHING

IX.1 Introduction

This chapter describes the use of the vector representation discussed in preceding chapters as a strategy for document retrieval. The search method adopted here is a simple linear associative technique. The vector representation of the question formulated in natural language is compared with each of the document vectors in the collection, and a measure of similarity for each question/document pair is computed. The documents are then ranked and retrieved in descending order of this similarity coefficient. Assuming that similarity corresponds to relevance, the documents are retrieved in order of relevance to the question, the most relevant being retrieved first.

Other examples of search systems using a vector technique are the UNIDATA system of Vischer (68) and the SMART system of Salton (11). A program incorporating vector searching - the QUANTUM program - was developed in 1971 by Niblett (69) to estimate the damages to be awarded to a plaintiff in a personal injuries action.

In the following two sections we compare vector searching with the traditional Boolean method, in which the question is formulated as keywords combined by the logical

operators AND, OR and NOT. The STATUS system developed at Harwell, and described in references (70) and (71), employs a Boolean search technique. An introduction to the properties of Boolean operators is given by Becker and Hayes (72).

IX.2 Searching as a classificatory process

We can view document searching as a form of classifying these documents with respect to the question asked. In a Boolean search the documents are classified into two groups, those which satisfy the Boolean expression and those which do not.

The vector search reaches a similarity coefficient of zero at some point in the ranking of documents, so that the collection can also be divided in two in this case. One group contains documents having a positive value of similarity with the question. The other documents have zero similarity. However, from the additional information contained in the ranking of documents we can construct a hierarchy of the documents in the first group, and we can make this group more homogeneous by choosing some threshold greater than zero.

In the following section the set of documents which satisfy the question is called the 'relevant class', the set of remaining documents the 'non-relevant class'.

IX.3 Vector versus Boolean searching

Vector searching possesses several advantages over Boolean methods. The relevant class obtained from a vector search is in most cases much larger; that is, recall is higher. This is because it contains documents which only partially satisfy the question, though of course these are ranked lower than those which completely satisfy it. On the other hand, documents are included in the equivalent Boolean relevant class if and only if they satisfy the Boolean expression completely. Documents which only partially satisfy it, but which might nevertheless be relevant, are rejected. To retrieve such documents a new Boolean question has to be formulated, which is less strict than the one which excluded the documents.

It can be a long and tedious process to find all the documents which partially satisfy the original Boolean question, since the term, or terms, in the Boolean expression which these documents lack may be different for each document. A new expression excluding the appropriate terms must be used to retrieve each document. In contrast, one vector question finds all the documents which are completely or partially relevant. Because these are retrieved at one go, the ranking is necessary to distinguish the highly relevant documents from the marginally relevant.

In reference (73) Lancaster has investigated the reasons for poor retrieval performance. One is the use of a wrong level of searching; the question is either too specific or too general for the level of indexing of the documents.

In considering full-text systems, we can regard the documents as being indexed 100% exhaustively; the full range of subject matter in each document is indexed. The specificity is also 100%, that is, the index terms are as specific as those used in the text, since they are indeed the same words. Thus we need to formulate questions both exhaustively and specifically. This can best be done by the vector method where the questions are expressed in natural language, which matches the level of indexing.

The Boolean method is less appropriate in this case. It requires the searcher to convert his question , which he originally thinks of in natural language, into a Boolean expression using terms which he thinks adequately describe his question. This he does at a certain level of specificity and exhaustivity of which he is probably unaware, and which is unlikely to match up to the indexing policy. In any case, to intentionally formulate a Boolean question which is highly exhaustive and specific will probably result in very low recall, as few documents will completely satisfy such an expression.

Another advantage of natural language questions is their ease of use, and vector searching is particularly attractive to people with little experience of indexing systems and Boolean algebra.

IX.4 The search program

This program accepts as input a question in natural language and, using the vector search technique, provides references to documents which are relevant to the question, in order of their relevance.

IX.4.1 Data required

Several files of data are required by the search program. Some of these are created by the programs described in Chapter IV.

(a) Dictionary arrays

A preliminary program converts the tree array and hash table created by the dictionary program (see section IV.2.3) into a form suitable for use in the search program. The tree array is converted to a direct access file, by dividing the original file into suitably-sized blocks. The file is held on disc, and the appropriate block read into core as required.

A new hash table is created. Each location, accessed by the hash value for some word, contains two fields; one is the number of the block in the tree file where the record for the corresponding word is to be found; the other is the location of the start of the record within that block. This hash table is held entirely in core. Figure IX.4 shows the relation between the hash table and tree file.

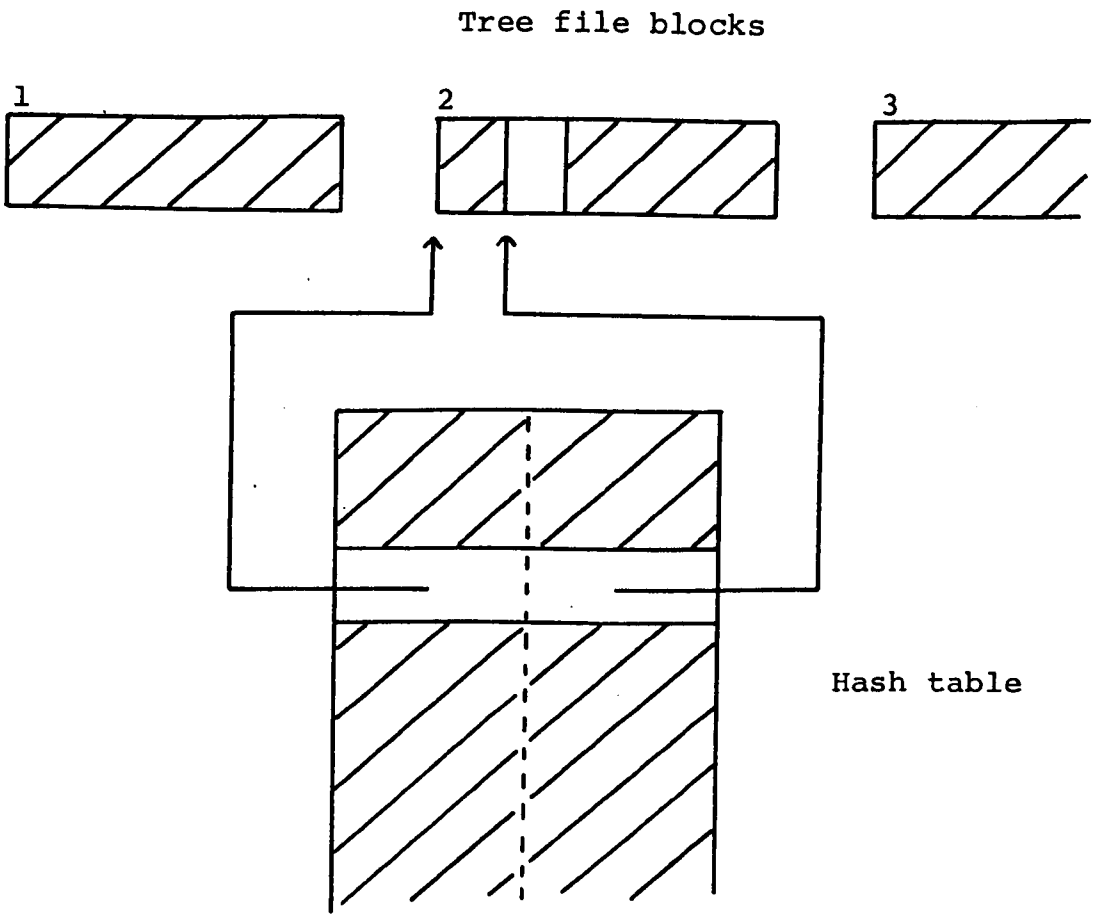


FIGURE IX.4 Tree file and hash table arrangement

(b) Concordance

The concordance file, arranged in blocks, and the block pointer array for this file (see section IV.7.2) are required by the search program. Concordance blocks are read in from disc as required, and the block pointer array is held in core.

(c) Vectors

The document vectors for the collection to be searched are required, in the form described in section IV.3.1. These are stored on disc, and the appropriate ones read in as required for comparison with the question vector.

(d) Titles

A file of titles, one for each document in the collection, is constructed for use by the search program. Each record in this file is 80 bytes long, accommodating a maximum of 80 characters for each title. The titles are stored in a direct access file on disc. The ith record in this file is the title of the document whose reference number is i. The relevant titles are retrieved from disc in answer to the question.

IX.4.2 Action of the program

A question in natural language is input to the program. This question must not exceed some maximum number of characters in length, this number being specified by the

program. An error message is returned if the question is too long. The maximum number of characters is usually set at 500, which is quite adequate for most questions.

A vector representation of the question is constructed in a similar manner to that used for document vectors in section IV.3. Words which occur in the question but which are not entered in the dictionary are not included in the question vector. Such words would contribute only to the denominator of the coefficient of similarity between the question and documents, and low coefficients would result if many of these words were included in the question vector. Common words are excluded from the document vectors for searching, and so too from the question vector.

Each word number in the question vector is then used to access the concordance block pointer array, which in turn enables the appropriate concordance block to be read into core. The references to the documents in which the corresponding word occurs are read from the concordance block. The lists of references for each word number in the question vector are merged to give one list containing each relevant document number once only. These are all the documents which contain at least one of the words in the question. The remaining documents in the collection will have zero coefficient of similarity with the question as they have no words in common with it.

The next step calculates the coefficient of similarity between the question vector and each of the relevant document vectors. These coefficients are sorted into descending numerical order, and the titles of the documents corresponding to the coefficients are retrieved in the specified order. The result of the search consists of a printout of these titles in decreasing order of relevance according to their coefficients of similarity.

IX.4.3 Search results

Certain constraints may be applied to the set of documents to be retrieved by the vector search. We may wish to specify a certain cut-off point, other than zero, for the similarity coefficient, so that only those documents having a coefficient greater than this value are retrieved. Alternatively, it is possible to specify that no more than some fixed number of documents are required, say ten, so that just the first ten documents in the ranking are retrieved.

These two conditions can also operate simultaneously. For example, the first ten documents having a coefficient greater than 0.5 could be specified. If less than ten documents have a sufficiently large coefficient, then less than ten documents are retrieved. Ten is the maximum number that can be retrieved, even if there are more with coefficients greater than 0.5. That is to say, both

conditions must always be satisfied.

The number of coefficients to be calculated is not reduced by imposing these constraints on the results, as it is not known beforehand which will satisfy the specified conditions. However, these constraints do serve to prevent the searcher being overwhelmed with output, since the system presents just those documents which are likely to be of most interest. Savings can also be made on storage space, as only the coefficients which satisfy the conditions need be reserved.

IX.5 Test run of search program

So far the search program has been used for a few small-scale tests with the Council of Europe data, which was converted to vector form for the experiments described in Chapter V.

The program is available in batch mode only at present, and no facility is yet included for printing out the text of the relevant documents. The titles are retrieved and printed copies of the documents must be consulted for the text. However, the results of the trial searches appeared promising, and it is hoped that the program will be elaborated on in the future. When this vector search program is fully operational, it will be possible to compare its performance with the Boolean system,

STATUS, operating in identical circumstances; that is, with the same machine and the same data bases.

IX.6 Further suggestions for vector searching

IX.6.1 Searching on word pairs

In section V.6 we used the frequencies of occurrence of word pairs as vector elements, but found that these performed less well than single words for classification. It was suggested that word pairs might more usefully be incorporated into a retrieval system.

To search on word pairs would require a further file to be available to the program, namely the word pair vectors for the document collection. A word pair vector would also have to be constructed for the question. The searcher would then have the option of searching on single words only, word pairs only, or on some combination of the two.

Single words and word pairs could be used in combination as follows. Two lists of relevant documents would first be produced for searches on single words and word pairs individually. These would then be merged to give a single list of relevant documents, whose coefficients of similarity with the question are given by the expression

$$\alpha\rho_1 + (1-\alpha)\rho_2$$

where ρ_1 and ρ_2 are the original coefficients for the single words and word pairs searches respectively. Some documents in the list for the single word search might not also be retrieved by the word pair search, and in this case the value of ρ_2 would be taken as zero. The term α in the expression is some number between 0 and 1, and would be specified by the searcher to allow due weight to be given to either single words or word pairs, whichever were considered the more important. Finally, the relevant documents would be sorted in order of the new combined coefficient, and the results presented in the usual way.

IX.6.2 Term weighting

In section II.2.2 we discussed the reasons for choosing document frequencies of words as vector elements, and referred to Sparck Jones' discussion (24) of other weightings for searching purposes. As the vectors constructed from document frequencies were readily available these were used for testing the search program. However, further investigations in vector searching might consider using document vectors whose elements measure some other weighting function applied to the terms in the documents.

CHAPTER X

CONCLUSION

X.1 Summary

The aim of this thesis has been to study the automatic processing of legal documents. Two areas have been examined; classification of documents, which has formed the major part of the thesis, and the searching of the full text of documents.

X.1.1 Classification of legal documents

Single-link cluster analysis has been used to classify several document collections, making use of various characteristics of the contents of the documents. The first and largest experiment used the full text of the Treaties of the Council of Europe. Another application of clustering examined the use of the common or function words extracted from the text of some statements used in criminal proceedings in order to determine authorship. The third investigation used the marginal citations of Statutes to produce a classification scheme of such documents.

The main conclusions from these experiments are as follows:

(1) The frequencies of occurrence of words in the full-text of the Treaties provide a sound basis for their classification. Excluding the very common, evenly distributed words gives a better classification than does the complete text.

(2) A whole Treaty is too large a body of text to be considered as a single document, since it is a non-homogeneous collection of formal clauses of signature, ratification etc., and non-formal clauses containing the substance of the Treaty, and these are best considered individually. A collection of articles clustered well into formal and non-formal groups, the non-formal clusters being formed on a subject basis.

(3) Single words perform better than word pairs as document descriptors. Word pair representations of documents are too specific so that the coefficients of similarity between such pairs of representatives are low and clustering is poor.

(4) Clustering on the basis of the common words only can distinguish varying styles of text.

(5) The occurrences of marginal citations can be used as a basis for grouping semantically related Acts of Parliament.

X.1.2 Searching legal documents

Vector searching has been implemented for the full text of the Treaties of the Council of Europe. Documents are retrieved according to their value of similarity with the question, calculated by the cosine coefficient. As yet no firm conclusions can be drawn as to its effectiveness as a search method, since large scale tests have not yet been carried out, nor have any of the possible sophistications, which might improve the performance, been implemented.

X.2 Recommendations

X.2.1 Classification

The exclusion of the common words, which tend to disrupt the classification, was found to improve the clustering based on full-text. It is therefore suggested that more rigorous ways of selecting the common words might be used, in order to improve still further on the results presented in Chapter V, where the common words were chosen subjectively.

Single-link was the only clustering method used in this study. It has considerable mathematical advantages over most other methods, and is simple to implement. However, for particular purposes some other representation of the

information contained in the matrix of similarity coefficients might be more useful. In particular, the investigation of overlapping clusters is suggested. These allow a document to be clustered with several groups according to the different topics it contains.

Another possible means of representation would be to construct a network of documents, similar to the term network devised by Doyle (74), in which a link is indicated between each pair having a coefficient of similarity greater than some specified value. This arrangement is easy to update, as the inclusion of a new document does not disrupt the network, as it might a clustering scheme. It is necessary only to link the new document to those with which it is sufficiently highly associated.

Determining style and authorship is another interesting and promising application of cluster analysis. The occurrences of common words were used as vector elements in the investigation in Chapter VII. However, other features of text have been used in traditional studies of style, for example the number of nouns and verbs used and sentence length, and these could all be correlated simultaneously using cluster analysis to give an overall view of the text.

The results of clustering on the basis of marginal citations were encouraging. As was suggested in Chapter VIII, there is scope for research into the automatic extraction of

citations from marginal notes, which requires an examination of the full-text in some cases, to complete or correct erroneous citations.

X.2.2 Searching

The SMART vector search system described in reference (11) has been extensively evaluated. However, whereas this system constructs vectors from the occurrences of index terms, the search system described in Chapter IX uses full-text. Therefore the performance of vector searching, shown to be effective in the SMART system, needs to be evaluated under full-text conditions.

A major feature to examine is the cost of searching; much more computing is involved in comparing the lengthy full-text vectors, than for comparing index term vectors or for conducting a Boolean search. Another important factor is the response time. In an interactive mode, search results need to be presented rapidly, as users are reluctant to sit idle at a terminal for any length of time.

Another possible future development in searching is the combination of the vector and Boolean methods. The results of a general Boolean search, yielding a large number of relevant documents, could be ranked by the vector method before presentation to the user. This would be the approach

when, for example, it was absolutely necessary that a certain word or words occurred in the documents retrieved. The vector search does not guarantee that any particular word in the question occurs in all the retrieved documents, hence a preliminary Boolean search would be required.

Finally, marginal citations could be considered for inclusion in a system searching Statutes. They could be used as document descriptors and search terms in addition to the words of full text. Alternatively, a classification scheme based on citations could be used to supplement search results, or for any of the other functions discussed in section I.4.

BIBLIOGRAPHY

(1) Myers, J. M. Computers and the searching of law texts in England and North America: A review of the state of the art Journal of Documentation 29 (2) 1973, 212 - 228

(2) Tapper, c. British experience in legal information retrieval Modern Uses of Logic in Law December 1964, 127 - 134

(3) Borko, H. Measuring the reliability of subject classification by men and machines American Documentation 15 1964, 268 - 273

(4) Keen, E. M. The Aberystwyth index languages test Journal of Documentation 29 (1) 1973, 1 - 35

(5) Sayers, W. C. B. An Introduction to Library Classification. 9th ed. Grafton and Co.; London. 1955.

(6) Sokal, R. R. Numerical taxonomy
Scientific American 215 (6)
1966, 106 - 116

(7) Maron, M. E. Automatic indexing: an
experimental inquiry
Journal of the Association for
Computing Machinery 8 (3)
1961, 404 - 417

(8) Borko, H.
and
Bernick, M. Automatic document
classification
Journal of the Association for
Computing Machinery 10 (2)
1963, 151 - 162

(9) Borko, H.
and
Bernick, M. Automatic document
classification Part II
Journal of the Association for
Computing Machinery 11 (2)
1964, 138 - 151

(10) Hoyle, W. G. Automatic indexing and
generation of classification
systems by algorithm
Information Storage and
Retrieval 9 (4) 1973, 233 - 242

(11) Salton, G. — The SMART Retrieval System - Experiments in Automatic Document Processing. Prentice-Hall; New Jersey. 1971.

(12) Dattola, R. T. — Experiments with a fast algorithm for automatic classification in reference (11), pp. 265 - 297

(13) Worona, S. — Query clustering in a large document space in reference (11), pp. 298 - 310

(14) Salton, G. — Some experiments in the generation of word and document associations Proceedings of the Fall Joint Computer Conference 22, AFIPS, 1962. pp. 234 - 250

(15) Sparck Jones, K. and Jackson, D. — Current approaches to classification and clump-finding at the Cambridge Language Research Unit Computer Journal 10 1967, 29 - 37

(16) van Rijsbergen, C. J. — Automatic Information Structuring and Retrieval. Ph.D. Thesis; University of Cambridge. 1972.

(17) Prywes, N. S. and Smith, D. P. — Organization of information in, Annual Review of Information Science and Technology 7 ed. C. Cuadra. ASIS; Washington DC. 1972. p. 149

(18) Salton, G. — Recent studies in automatic text analysis and document retrieval Journal of the Association for Computing Machinery 20 (2) 1973, 258 - 278

(19) Jardine, N. and van Rijsbergen, C. J. — The use of hierarchic clustering in information retrieval Information Storage and Retrieval 7 (5) 1971, 217 - 240

(20) Prywes, N. S. and Litofsky, B. — All automatic processing for a large library Proceedings of the Spring Joint Computer Conference 36, AFIPS; 1970. pp. 323 - 331

(21) Swanson, D. R. — Searching natural language by computer. Science 132 (3434) 1960, 1099 - 1104

(22) Sparck Jones, K. — Some thoughts on classification for retrieval. Journal of Documentation 26 (2) 1970, 89 - 101

(23) Patterson, E. M. and Rutherford, D. E. — Elementary Abstract Algebra. University Mathematical Texts, Oliver and Boyd Ltd; London. 1965.

(24) Sparck Jones, K. — Automatic Indexing - A State of the Art Review. Computer Laboratory; University of Cambridge. 1974.

(25) Sparck Jones, K. — A statistical interpretation of term specificty and its application in retrieval. Journal of Documentation 28 (1) 1972, 11 - 21

(26) Weiss, S. F. — Syntax in Text Analysis. Cornell University, Ithaca; New York. 1969. Information Storage and Retrieval Series. Scientific Report. ISR - 16.

(27) Weiss, S. F. — Template Analysis and its Application to Natural Language Processing. Cornell University, Ithaca; New York. 1969. Information Storage and Retrieval Series. Scientific Report. ISR - 16.

(28) Salton, G. — Automatic Information Organization and Retrieval. McGraw-Hill; New York. 1968. pp. 151 - 201

(29) Neufeld, M. L., Graham, K. L. and Mazella, A. — Machine-aided title word indexing for a weekly current awareness publication Information Storage and Retrieval 10 (11/12) 1974, 403 - 410

(30) Hill, D. R.

A vector clustering technique in, Mechanised Information Storage, Retrieval and Dissemination.
ed. K. Samuelson
Proceedings of the FID/IFIP Joint Conference, Rome, June 1967.
North-Holland Publishing Co.; Amsterdam. 1968. pp. 225 - 234

(31) Sneath, P. H. A.

The application of computers to taxonomy
Journal of General Microbiology 17 1957, 201 - 226

(32) Michener, C. D. and Sokal, R. R.

A quantitative approach to a problem in classification
Evolution 11 1957, 130 - 162

(33) Sneath, P. H. A. and Sokal, R. R.

Numerical Taxonomy.
W. H. Freeman and Co.; San Francisco. 1973.

(34) Ball, G. H. Data analysis in the social sciences: what about the details? Proceedings of the Fall Joint Computer Conference 27 part 1. AFIPS; 1965. pp. 533 - 559

(35) Cormack, R. M. A review of classification Journal of the Royal Statistical Society Series A 134 (3) 1971, 321 - 367

(36) Jardine, N. and Sibson, R. Mathematical Taxonomy. Wiley; London. 1971.

(37) Everitt, B. Cluster Analysis. Heinemann Educational Books Ltd; London. 1974.

(38) Beckner, M. The Biological Way of Thought. Columbia University Press; New York. 1959.

(39) Zadeh, L. A. Fuzzy sets Information and Control 8 (3) 1965, 338 - 353

(40) Williams, W. T. and Lambert, J. M. — Multivariate methods in plant ecology. I. Association analysis in plant communities Journal of Ecology 47 1959, 83 - 101

(41) Jardine, N. and Sibson, R. — The construction of hierarchic and non-hierarchic classifications Computer Journal 11 1968, 177 - 184

(42) Needham, R. M. and Sparck Jones, K. — Keywords and clumps Journal of Documentation 20 (1) 1964, 5 - 15

(43) Williams, W. T., Clifford, H. T. and Lance, G. N. — Group-size dependence: a rationale for choice between numerical classifications Computer Journal 14 (2) 1971, 157 - 162

(44) Williams, W. T., Lance, G. N., Dale, M. B. and Clifford, H. T. — Controversy concerning the criteria for taxonometric strategies Computer Journal 14 (2) 1971, 162 - 165

(45) Sibson, R.
Some observations on a paper by Lance and Williams
Computer Journal 14 (2)
1971, 156 - 157

(46) Jardine, N.
and
Sibson, R.
Choice of methods for automatic classification
Computer Journal 14 (4)
1971, 404 - 406

(47) van Rijsbergen, C. J.
A fast hierarchic clustering algorithm
Computer Journal 13 (3)
1970, 324 - 326

(48) Wishart, D.
A generalisation of nearest neighbour which reduces chaining effects
in, Numerical Taxonomy.
ed. A. J. Cole
Academic Press; New York. 1969.
pp. 282 - 311

(49) Sokal, R. R.
and
Sneath, P. H. A.
Principles of Numerical Taxonomy.
W. H. Freeman and Co.;
San Francisco. 1963.

(50) Goodall, D. W. — A new similarity index based on probability. Biometrics 22 (4) 1966, 882 - 907

(51) Rubin, J. — Optimal Classification into Groups, an Approach for Solving the Taxonomy Problem. IBM New York Scientific Centre; New York. March 1966.

(52) Sokal, R. R. — The comparison of dendrograms by objective methods. Taxon 11 (2) 1962, 33 - 40

(53) Day, A. C. — FORTRAN Techniques (with special reference to non-numerical applications). Cambridge University Press; London. 1972. pp. 76 - 81

(54) Lum, V. Y., Yuen, P. S. T. and Dodd, M. — Key-to-address transform techniques: a fundamental performance study on large existing formatted files. Communications of the Association for Computing Machinery 14 (4) 1971, 228 - 239

(55) Lowe, T. C. Effectiveness of retrieval key abbreviation schemes
Journal of the American Society for Information Science 22 (6) 1971, 374 - 381

(56) Council of Europe Chart showing Signatures and Ratifications of Council of Europe Conventions and Agreements.
Directorate of Legal Affairs; Council of Europe. February, 1974.

(57) Good, I. J. Statistics of language: introduction
in, Encyclopaedia of Linguistics, Information and Control.
Ed. A. R. Meetham
Pergamon Press; Oxford. 1969.
pp. 567 - 581

(58) Zipf, G. K. Human Behaviour and the Principle of Least Effort.
Addison-Wesley; Cambridge, Massachusetts. 1949.

(59) Ule, L. A. Cluster analysis
Association for Literary and Linguistic Computing Bulletin
2 (3) 1974, 16 - 21

(60) Ule, L. A. — The use of CONSTAT in authorship investigations. Association for Literary and Linguistic Computing Bulletin 3 (3) 1975, 211 - 225

(61) Radday, Y. T. and Shore, H. — The definite article: a type- and/or author-specifying discriminant in the Hebrew Bible. Association for Literary and Linguistic Computing Bulletin 4 (1) 1976, 23 - 31

(62) Morton, A. Q. and Winspear, A. D. — It's Greek to the Computer. Harvest House; Montreal. 1971.

(63) Mosteller, F. and Wallace, D. L. — Inference and Disputed Authorship: the Federalist. Addison-Wesley Publishing Co.; Reading, Massachusetts. 1964.

(64) Price, N. and Schiminovich, S. — A clustering experiment: first step towards a computer generated classification scheme Information Storage and Retrieval 4 (3) 1968, 271 - 280

(65) Michelson, D., Amreich, M., Grissom, G. and Ide, E. — An experiment in the use of bibliographic data as a source of relevance feedback in information retrieval in reference (11), pp. 430 - 443

(66) Salton, G. — Associative document retrieval techniques using bibliographic information Journal of the Association for Computing Machinery 10 (4) 1963, 440 - 457

(67) de Solla Price, D. J. — Networks of scientific papers Science 149 (3683) 1965, 510 - 515

(68) Vischer, B. — Das Dokumentationssytem der UNIDATA AG Kybernetik. Daterverarbeitung. Recht, Materialien zur Rechtsinformatik 1 1971, 71 - 87

(69) Niblett, B.

Private communication.

(70) Niblett, B. and Price, N. H.

Mechanised searching of Acts of Parliament
Information Storage and Retrieval 6 (3) 1970, 289 - 297

(71) Price, N. H., Bye, C. and Niblett, B.

On-line searching of Council of Europe Conventions and Agreements: A study in bilingual document retrieval
Information Storage and Retrieval 10 (3/4) 1974, 145-154

(72) Becker, J. and Hayes, R. M.

Information Storage and Retrieval: tools, elements, theories.
Wiley; New York. 1963.

(73) Lancaster, F. W.

Information Retrieval Systems: Characteristics, Testing and Evaluation.
Wiley; New York. 1968

(74) Doyle, L. B.

Semantic road maps for literature searchers
Journal of the Association for Computing Machinery 8 (4)
1961, 553 - 578

APPENDIX A

MATHEMATICAL DEFINITIONS FOR VECTOR THEORY

In the following definitions the symbol 'ε' means 'is a member of '; e.g. $x \varepsilon S$ means that x is a member of the set S.

1. Groups

Definition: An Abelian Group is a set G with a binary operation * defined on it such that the following conditions hold:

(1) G is closed with respect to *

i.e. For all $g, h \varepsilon G$

$$g * h \varepsilon G$$

(2) * is associative

i.e. For all $f, g, h \varepsilon G$

$$f * (g * h) = (f * g) * h$$

(3) * is commutative

i.e. For all $g, h \varepsilon G$

$$g * h = h * g$$

(4) G contains an identity element with respect to *

i.e. There exists $e \varepsilon G$ such that, for all $g \varepsilon G$

$$g * e = e * g = g$$

(5) every element of G has an inverse in G with respect to *

i.e. For all $g \in G$, there exists $g' \in G$ such that $g * g' = g' * g = e$

Conditions (1), (2), (4) and (5) define any group. Condition (3) is necessary for an Abelian, i.e. commutative, group.

An example of an Abelian group is the set of integers (positive, negative and zero), with the binary operation of addition. Zero is the identity element, and the inverse of any integer x is -x.

2. Fields

Definition: A Field is a set F having at least two elements, with two binary operations + (addition) and * (multiplication) defined on it such that the following conditions hold:

(1) F is an Abelian group with respect to +

(2) the non-zero elements of F form an Abelian group with respect to * (where zero is the identity element with respect to +)

(3) the distributive laws hold in F

i.e. For all $x, y, z \in F$

$$x * (y + z) = x * y + x * z$$

and $(x + y) * z = x * z + y * z$

The set of all real numbers with the usual operations of addition and multiplication is an example of a field.

3. Vector Spaces

Definition: A Vector Space V over a field F is a set with two binary operations + (addition) and * (scalar multiplication by F) defined on it such that the following conditions hold:

(1) V is an Abelian group with respect to +

(2) V is closed with respect to multiplication by F
i.e. For all $v \in V$, $a \in F$
$a * v \in V$

(3) the distributive laws hold in V for multiplication by F
i.e. For all $v \in V$, $a, b \in F$
$(a + b) * v = a * v + b * v$
and for all $v, w \in V$, $a \in F$
$a * (v + w) = a * v + a * w$

(4) multiplication by F is associative
i.e. For all $a, b \in F$, $v \in V$
$(a * b) * v = a * (b * v)$

(5) if 1 is the identity element in F, i.e. for all $a \in F$ $1 * a = a * 1 = a$
then for all $v \in V$
$1 * v = v$

The set of complex numbers is a vector space over the field of real numbers.

APPENDIX B

TITLES OF THE TREATIES OF THE COUNCIL OF EUROPE

1. Statute of the Council of Europe. (1)

2. Convention for the Protection of Human Rights and Fundamental Freedoms, and Protocol. (5 and 9)

3. European Interim Agreement on Social Security Schemes relating to Old Age, Invalidity and Survivors, and Protocol thereto. (12)

4. European Interim Agreement on Social Security other than Schemes for Old Age, Invalidity and Survivors, and Protocol thereto. (13)

5. European Convention on Social and Medical Assistance, with Annexes and Protocol. (14)

6. European Convention on the Equivalence of Diplomas leading to Admission to Universities. (15)

7. European Convention Relating to Formalities required for Patent Applications. (16)

8. European Convention on the International Classification of Patents for Invention. (17)

9. European Cultural Convention. (18)

10. European Convention on Establishment, and Protocol. (19)

11. Agreement on the Exchange of War Cripples between Member Countries of the Council of Europe with a view to Medical Treatment. (20)

12. European Convention on the Equivalence of Periods of University Study. (21)

13. Second Protocol to the General Agreement on Privileges and Immunities of the Council of Europe. (22)

14. European Convention for the Peaceful Settlement of Disputes. (23)

15. European Convention on Extradition. (24)

16. European Agreement on Regulations governing the Movement of Persons between Member States of the Council of Europe. (25)

17. European Agreement Concerning Programme Exchange by means of Television Films. (27)

18. European Convention on Compulsory Insurance against Civil Liability in respect of Motor Vehicles, with Annexes and Protocol. (29)

19. European Convention on Mutual Assistance in Criminal Matters. (30)

20. European Convention on the Academic Recognition of University Qualifications. (32)

21. Agreement on the Temporary Importation, Free of Duty, of Medical, Surgical and Laboratory Equipment for Use on Free Loan in Hospitals and other Medical Institutions for Purposes of Diagnosis and Treatment. (33)

22. European Agreement on the Protection of Television Broadcasts. (34)

23. European Social Charter, with Appendix. (35)

24. Fourth Protocol to the General Agreement on Privileges and Immunities of the Council of Europe Provisions concerning the European Court of Human Rights. (36)

25. European Agreement on Mutual Assistance in the Matter of Special Medical Treatments and Climatic Facilities. (38)

26. Convention on the Liability of Hotel-Keepers concerning the Property of their Guests, with Annex. (41)

27. Agreement Relating to Application of the European Convention on International Commercial Arbitration. (42)

28. Convention on Reduction of Cases of Multiple Nationality and on Military Obligations in Cases of Multiple Nationality, with Annex. (without reservations) (43)

29. Protocol No.2 of the Convention on the Protection of Human Rights and Fundamental Freedoms, conferring upon the European Court of Human Rights competence to give advisory opinions. (44)

30. Protocol No.3 to the Convention on the Protection of Human Rights and Fundamental Freedoms, amending Articles 29, 30 and 34 of the Convention. (45)

31. Protocol No.4 to the Convention for the Protection of Human Rights and Fundamental Freedoms, securing certain rights and freedoms other than those already included in the Convention and in the first Protocol thereto. (46)

32. Convention on the Unification of Certain Points of Substantive Law on Patents for Invention. (47)

33. Protocol to the European Convention on the Equivalence of Diplomas leading to Admission to Universities. (49)

34. Convention on the Elaboration of a European Pharmacopoeia. (50)

35. European Convention on the Supervision of conditionally sentenced or conditionally released Offenders, with Annex. (51)

36. European Convention on the Punishment of Road Traffic Offences, with Annexes. (52)

37. European Agreement for the Prevention of Broadcasts transmitted from Stations outside National Territories. (53)

38. Protocol to the European Agreement on the Protection of Television Broadcasts. (54)

39. Protocol No.5 to the Convention for the Protection of Human Rights and Fundamental Freedoms, amending Articles 22 and 40 of the Convention. (55)

40. European Convention on Establishment of Companies, and Protocol. (57)

41. European Convention on the Adoption of Children. (58)

42. European Agreement on the Instruction and Education of Nurses, with Annexes. (59)

43. European Convention on Foreign Money Liabilities, with Annex. (60)

44. European Convention on Information on Foreign Law. (62)

45. European Convention on the Abolition of Legislation of Documents executed by Diplomatic Agents or Consular Officers. (63)

46. European Convention on the restriction of the use of certain Detergents in washing and cleansing Products. (64)

47. European Convention for the Protection of Animals during International Transport. (65)

48. European Convention on the Protection of the Archaeological Heritage. (66)

49. European Agreement relating to persons participating in proceedings of the European Commission and Court of Human Rights. (67)

50. European Agreement on "au pair" Placement, with Annexes and Protocol. (68)

51. European Agreement on continued Payment of Scholarships to Students studying abroad. (69)

52. European Convention on the Repatriation of Minors. (71)

53. Convention relating to Stops on Bearer Securities in International Circulation, with regulations. (72)

54. Special Agreement relating to the Seat of the Council of Europe. (3)

55. Interpretation of the terms: "Nationals" and "Territory" in the interim agreements on Social Security and in the Convention on Social and Medical Assistance. (12, 13 and 14, modifications)

56. Annexes to the European Convention on Social and Medical Assistance and Protocol. (14, modifications)

57. Annexes to the European Interim Agreement on Social Security Schemes relating to Old Age, Invalidity and Survivors and Protocol thereto. (12, modifications)

58. Annexes to the European Interim Agreement on Social Security other than Schemes for Old Age, Invalidity and Survivors and Protocol thereto. (13, modifications)

59. European Convention on Extradition; Declarations and Reservations. (24, modifications)

60. European Convention providing a uniform Law on Arbitration, with Annex. (56)

61. European Convention on the International Validity of Criminal Judgements, with Appendices. (70)

62. European Agreement on Travel by Young Persons on Collective Passports between the Member Countries of the Council of Europe. (37)

63. European Agreement on the Exchange of Therapeutic Substances of Human Origin, with Protocol and Annexes. (26)

64. European Agreement on the Exchanges of Blood Grouping Reagents, with Protocol and Annex. (39)

65. Agreement between the Member States of the Council of Europe on the issue to Military and Civilian War-disabled of an International Book of Vouchers for the Repair of Prosthetic and Orthopaedic Appliances, with Annex. (40)

66. European Convention on Consular Functions
    (i) Protocol concerning the Protection of Refugees
    (ii) Protocol in respect of Civil Aircraft. (61)

67. European Code of Social Security and Protocol to the European Code of Social Security, with Annex and Addenda. (48)

68. European Convention on the Establishment of a Scheme of Registration of Wills. (77)

69. European Convention on Civil Liability for Damage by Motor Vehicles. (79)

70. European Convention on the Place of Payment of Money Liabilities. (75)

71. European Agreement on the Abolition of Visas for Refugees. (31)

72. European Convention on the Calculation of Time Limits. (76)

73. European Convention on the Transfer of Proceedings in Criminal Matters. (73)

74. European Convention on State Immunity. (74)

75. Additional Protocol to the European Convention on State Immunity. (74)

76. General Agreement on Privileges and Immunities of the Council of Europe. (2)

77. Supplementary Agreement to the General Agreement on Privileges and Immunities of the Council of Europe. (4)

78. Protocol to the General Agreement on Privileges and Immunities of the Council of Europe. (10)

79. Third Protocol to the General Agreement on Privileges and Immunities of the Council of Europe. (28)

80. Articles of Agreement of the Council of Europe Resettlement Fund. (28)

81. European Convention on Social Security. (78)

82. Supplementary Agreement for the Application of the European Convention on Social Security. (78)

83. Agreement on the Transfer of Corpses. (80)

84. Additional Protocol to the Protocol to the European Convention on the Protection of Television Broadcasts. (81)

85. European Convention on the non-applicability of Statutory Limitation to Crimes against Humanity and War Crimes. (82)

86. European Convention on the Social Protection of Farmers. (83)

APPENDIX C

ENGLISH TEXT OF THE TREATIES FORMING THE CLUSTER 'UNIVERSITY STUDY'

(The underlined number preceding each portion of text is the document number used as a reference in the clustering of articles in Chapter V.)

Treaty 6 European Convention on the Equivalence of Diplomas Leading to Admission to Universities

1. European Convention on the Equivalence of Diplomas Leading to Admission to Universities
The governments signatory hereto, being members of the Council of Europe,
considering that one of the objects of the Council of Europe is to pursue a policy of common action in cultural and scientific matters;
considering that this object would be furthered by making the intellectual resources of members freely available to European youth;
considering that the university constitutes one of the principal sources of the intellectual activity of a country;
considering that students who have successfully completed their secondary school education in the territory of one member should be afforded all possible facilities to enter a university of their choice in the territory of other members;
considering that such facilities, which are also desirable in the interests of freedom of movement from country to country, require the equivalence of diplomas leading to admission to universities,
have agreed as follows:

2. Article 1

1. Each contracting party shall recognise for the purpose of admission to the universities situated in its territory, admission to which is subject to state control, the equivalence of those diplomas awarded in the territory of each other contracting party which constitute a requisite qualification for admission to similar institutions in the country in which these diplomas were awarded.
2. Admission to any university shall be subject to the availability of places.
3. Each contracting party reserves the right not to apply the provisions of paragraph 1 to its own nationals.
4. Where admission to universities situated in the territory of a contracting party is outside the control of the state, that contracting party shall transmit the text of this Convention to the universities concerned and use its best endeavours to obtain the acceptance by the latter of the principles stated in the preceding paragraphs.

3. Article 2

Each contracting party shall, within a year of the coming into force of this Convention, provide the Secretary-General of the Council of Europe with a written statement of the measures taken to implement the previous article.

4. Article 3

The Secretary-General of the Council of Europe shall communicate to the other contracting parties the information received from each of them in accordance with Article 2 above and shall keep the committee of ministers informed of the progress made in the implementation of this Convention.

5. Article 4

For the purpose of this Convention:
(a) the term 'diploma' shall mean any diploma, certificate or other qualification, in whatever form it may be awarded or recorded, which entitles the holder or the person concerned to apply for admission to a university:

(b) the term 'universities' shall mean:
(i) universities;
(ii) institutions regarded as being similar in character to universities by the contracting party in whose territory they are situated.

6. Article 5

1. This Convention shall be open to the signature of the members of the Council of Europe. It shall be ratified. The instruments of ratification shall be deposited with the Secretary-General of the Council of Europe.
2. The Convention shall come into force as soon as three instruments of ratification have been deposited.
3. As regards any signatory ratifying subsequently, the Convention shall come into force at the date of the deposit of its instrument of ratification.
4. The Secretary-General of the Council of Europe shall notify all the members of the Council of Europe of the entry into force of the Convention, the names of the contracting parties which have ratified it, and the deposit of all instruments of ratification which may be effected subsequently.

7. Article 6

The committee of ministers of the Council of Europe may invite any state which is not a member of the Council to accede to this Convention. Any state so invited may accede by depositing its instrument of accession with the Secretary-General of the Council, who shall notify all the contracting parties thereof. As regards any acceding state, this Convention shall come into force on the date of the deposit of its instrument of accession.

8. In witness whereof the undersigned, being duly authorised thereto, have signed the present Convention. Done at Paris, this 11th day of December, 1953, in English and French, both texts being equally authoritative, in a single copy which shall remain deposited in the archives of the Council of Europe. The Secretary-General shall transmit certified copies to each of the signatories.

9. Declarations

Territorial application
Belgium
(Declaration made by the Minister for Foreign Affairs of Belgium, dated 21st May 1955)

Translation
On depositing the instrument of ratification of His Majesty the King of the Belgians relating to the European Convention on the Equivalence of Diplomas Leading to Admission to Universities, signed at Paris on 11th December 1953, I declare that the said instruments of ratification are valid only for the Metropolitan Territory of Belgium and that the Belgian Congo and the trustee territory of Ruanda-Urundi are expressly excluded.

10. Federal Republic of Germany

(Extract from the proces-verbal of deposit dated 3rd March 1955 of the instrument of ratification)

Translation
The Federal Government declares that the European Convention on the Equivalence of Diplomas Leading to Admission to Universities applies equally to Land Berlin.

Treaty 12 European Convention on the Equivalence of Periods of University Study

11. European Convention on the Equivalence of Periods of University Study
The governments signatory hereto, being members of the Council of Europe, having regard to the European Convention on the Equivalence of Diplomas Leading to Admission to Universities, signed at Paris on the 11th December, 1953;
having regard to the European Cultural Convention, signed in Paris on the 19th December, 1954;
considering that an important contribution would be made to European understanding if a larger number of students, among others students of modern languages, could spend a period of study abroad and if examinations passed and courses taken by such students during the period of study could be recognised by the home university;

considering further that the recognition of periods of study spent abroad would contribute to the solution of the problem raised by the shortage of highly qualified scientists,
have agreed as follows:

12. Article 1

1. For the purposes of the present Convention, contracting parties shall be divided into categories according to whether the authority competent to deal with matters pertaining to equivalences in their territories is:
(a) the state;
(b) the university;
(c) the state or university as the case may be.
Each contracting party shall inform the Secretary-General of the Council of Europe which is the competent authority in its territory to deal with matters pertaining to equivalences.
2. The term 'universities' shall denote:
(a) universities, and
(b) institutions regarded as being similar in character to universities by the contracting party in whose territory they are situated.

13. Article 2

1. Contracting parties falling within category (a) of Article 1, paragraph 1, shall recognise a period of study spent by a student of modern languages in a university of another member country of the Council of Europe as equivalent to a similar period spent in his home university provided that the authorities of the first-mentioned university have issued to such a student a certificate attesting that he has completed the said period of study to their satisfaction.
2. The length of the period of study referred to in the preceding paragraph shall be determined by the competent authorities of the contracting party concerned.

14. Article 3

Contracting parties falling within category (a) of Article 1, paragraph 1, shall consider the means to be adopted in order to recognise a period of study spent in a university of another member country of the

Council of Europe by students of disciplines other than modern languages and especially by students of pure and applied sciences.

15. Article 4

Contracting parties falling within category (a) of Article 1, paragraph 1, shall endeavour to determine, by means of unilateral or bilateral arrangements, the conditions under which an examination passed or a course taken by a student during a period of study in a university of another member country of the Council of Europe may be considered as equivalent to a similar examination passed or a course taken by a student in his home university.

16. Article 5

Contracting parties falling within category (b) of Article 1, paragraph 1, shall transmit the text of the present Convention to the authorities of the universities situated in their territories and shall encourage the favourable consideration and application by them of the principles mentioned in Articles 2, 3 and 4 above.

17. Article 6

Contracting parties falling within category (c) of Article 1, paragraph 1, shall apply the provisions of Articles 2, 3 and 4 in respect of those universities for which the state is the competent authority in the matters dealt with in this Convention, and shall apply the provisions of Article 5 in respect of those universities which are themselves the competent authorities in these matters.

18. Article 7

Each contracting party shall, within a year of the coming into force of the present Convention, furnish the Secretary-General of the Council of Europe with a written statement on the measures taken to implement Articles 2, 3, 4, 5 and 6.

19. Article 8

The Secretary-General of the Council of Europe shall communicate to the other contracting parties the information received from each of them in accordance with Article 7 and shall keep the committee of ministers informed of the progress made in the implementation of this Convention.

20. Article 9

1. The present Convention shall be open to the signature of the members of the Council of Europe. It shall be ratified. The instruments of ratification shall be deposited with the Secretary-General of the Council of Europe.
2. The Convention shall come into force as soon as three instruments of ratification have been deposited.
3. As regards any signatory ratifying subsequently, the Convention shall come into force at the date of the deposit of its instrument of ratification.
4. The Secretary-General of the Council of Europe shall notify all the members of the Council of Europe of the entry into force of the Convention, the names of the contracting parties which have ratified it and the deposit of all instruments of ratification which may be effected subsequently.
5. Any contracting party may specify the territories to which the provisions of the present Convention shall apply by addressing to the Secretary-General of the Council of Europe a declaration which shall be communicated by the latter to all the other contracting parties.

21. Article 10

The committee of ministers of the Council of Europe may invite any state which is not a member of the Council to accede to the present Convention. Any state so invited may accede by depositing its instrument of accession with the Secretary-General of the Council, who shall notify all the contracting parties thereof. Any acceding state shall be considered a member country of the Council of Europe for the purpose of the present Convention. As regards any acceding state, the present Convention shall come into force on the date of the deposit of its instrument of accession.

22. In witness whereof the undersigned, duly authorised thereto by their respective governments, have signed the present Convention.
Done at Paris, this 15th day of December, 1956, in the English and French languages, both texts being equally authoritative, in a single copy which shall remain deposited in the archives of the Council of Europe. The Secretary-General shall transmit certified copies to each of the signatory and acceding governments.

23. Declarations

Made in accordance with Article 9, paragraph 5
Territorial application
Federal Republic of Germany
(Letter from the permanent representative to the Council of Europe dated 24th February 1965)

Translation
The European Convention on the Equivalence of Periods of University Study of 15th December 1956 shall also apply to the Land Berlin with effect from 8th December 1964, i.e. the date on which it entered into force for the Federal Republic of Germany.

Netherlands
(Extract from the instrument of ratification)
Translation
The instrument of ratification of the Kingdom of the Netherlands specifies that the Convention shall apply to the Kingdom in Europe.

United Kingdom
(Extract from the proces-verbal of deposit dated 18th September 1957 of the instrument of ratification)

First declaration - 18th September 1957
On depositing this day on behalf of the Government of the United Kingdom of Great Britain and Northern Ireland the instrument of ratification of the European Convention on the Equivalence of Periods of University Study which was signed at Paris on the 15th of December 1956 I am directed by Her Majesty's Principal Secretary of State for Foreign Affairs to inform you that, while the said instrument is in respect of the United Kingdom of Great Britain and Northern Ireland only, the Government of the United Kingdom of Great Britain and Northern Ireland interpret paragraph 5 of Article 9 as permitting them to extend the application of the said Convention at any time hereafter to any territory for whose international relations they are responsible.

Second declaration - 2nd January 1958
With reference to the declaration made on behalf of the Government of the United Kingdom of Great Britain and Northern Ireland at the time of the deposit of their instrument of ratification of the European Convention on the Equivalence of Periods of University Study signed at Paris on the 15th of December, 1956, concerning their interpretation of paragraph 5 of Article 9 of the said Convention, I have the honour to inform your Excellency of the application of the above-mentioned Convention to the Federation of Rhodesia and Nyasaland with effect from this day's date.

Treaty 20 European Convention on the Academic Recognition of University Qualifications

24. European Convention on the Academic Recognition of University Qualifications
The governments signatory hereto, being members of the Council of Europe,
having regard to the European Cultural Convention, signed in Paris on 19th December 1954;
having regard to the European Convention on the Equivalence of Diplomas Leading to Admission to Universities, signed in Paris on 11th December 1953;
having regard to the European Convention on the Equivalence of Periods of University Study, signed in Paris on 15th December 1956;
considering the desirability of supplementing those Conventions by providing for the academic recognition of university qualifications obtained abroad,
have agreed as follows:

25. Article 1

For the purpose of the present Convention:
(a) the term 'universities' shall denote
(i) universities, and
(ii) institutions regarded as being of university level by the contracting party in whose territory they are situated and having the right to confer qualifications of university level;
(b) the term 'university qualification' shall denote any degree, diploma or certificate awarded by a university situated in the territory of a contracting party and marking the completion of a period of university study;

(c) degrees, diplomas and certificates awarded on the results of a part-examination shall not be regarded as university qualifications within the meaning of sub-paragraph (b) of the present Article.

26. Article 2

1. For the purpose of the present Convention, contracting parties shall be divided into categories according to whether the authority competent in their territory to deal with matters pertaining to the equivalence of university qualifications is:
(a) the state;
(b) the university;
(c) the state or the university, as the case may be.
2. Each contracting party shall, within one year of the coming into force of the present Convention in respect of itself, inform the Secretary-General of the Council of Europe which is the authority competent in its territory to deal with matters pertaining to the equivalence of university qualifications.

27. Article 3

1. Contracting parties falling within category (a) in paragraph 1 of Article 2 of the present Convention shall grant academic recognition to university qualifications conferred by a university situated in the territory of another contracting party.
2. Such academic recognition of a foreign university qualification shall entitle the holder:
(a) to pursue further university studies and sit for academic examination on completion of such studies with a view to proceeding to a further degree, including that of a doctorate, on the same conditions as those applicable to nationals of the contracting party, where admission to such studies and examinations depends upon the possession of a similar national university qualification;
(b) to use an academic title conferred by a foreign university, accompanied by an indication of its origin.

28. Article 4

In respect of sub-paragraph 2(a) of Article 3 of the present Convention, each contracting party may:
(a) in cases where the examination requirements for a foreign university qualification do not include certain subjects prescribed for the similar national qualification, withhold recognition until a

supplementary examination has been passed in the subjects in question;
(b) require holders of a foreign university qualification to pass a test in its official language, or one of its official languages, in the event of their studies having been pursued in another language.

29. Article 5

Contracting parties falling within category (b) in paragraph 1 of Article 2 of the present Convention shall transmit the text of the Convention to the authorities competent in their territory to deal with matters pertaining to the equivalence of university qualifications and shall encourage the favourable consideration and application by them of the principles set out in Articles 3 and 4 thereof.

30. Article 6

Contracting parties falling within category (c) in paragraph 1 of Article 2 of the present Convention shall apply the provisions of Articles 3 and 4 thereof where the state is the authority competent to deal with the equivalence of university qualifications and shall apply the provisions of Article 5 thereof where the state is not the competent authority in these matters.

31. Article 7

The Secretary-General of the Council of Europe may from time to time request contracting parties to furnish a written statement on the measures and decisions taken with a view to implementing the provisions of the present Convention.

32. Article 8

The Secretary-General of the Council of Europe shall communicate to the other contracting parties the information received from each of them in accordance with Articles 2 and 7 of the present Convention and shall keep the committee of ministers informed of the progress made in the implementation of the present Convention.

33. Article 9

Nothing in the present Convention shall be deemed:
(a) to effect any more favourable provisions concerning the recognition of foreign university qualifications contained in an existing Convention to which a contracting party may be signatory or to render less desirable the conclusion of any further such Convention by any of the contracting parties, or
(b) to prejudice the obligation of any person to comply with the laws and regulations in force in the territory of any contracting party concerning the entry, residence and departure of foreigners.

34. Article 10

1. The present Convention shall be open to the signature of the members of the Council of Europe. It shall be ratified. The instruments of ratification shall be deposited with the Secretary-General of the Council of Europe.
2. The Convention shall enter into force one month after the date of deposit of the third instrument of ratification.
3. In respect of any signatory ratifying subsequently, the Convention shall enter into force one month after the date of deposit of its instrument of ratification.
4. After the entry into force of the present Convention, the committee of ministers may invite any state which is not a member of the Council to accede thereto. Any state so invited may accede by depositing its instrument of accession with the Secretary-General of the Council. As regards an acceding state, the present Convention shall enter into force one month after the date of deposit of its instrument of accession.
5. The Secretary-General of the Council of Europe shall notify all members of the Council and any acceding state of the deposit of all instruments of ratification and accession.

35. Article 11

Any contracting party may, at the time of deposit of its instrument of ratification or accession, or at any time thereafter, declare by notification addressed to the Secretary-General of the Council of Europe that the present Convention shall apply to some or all of the territories for the international relations of which it is responsible.

36. Article 12

1. Any contracting party may denounce the present Convention at any time after it has been in force for a period of five years by means of a notification addressed to the Secretary-General of the Council of Europe, who shall so inform the other contracting parties.
2. Such denunciation shall take effect in respect of the contracting party concerned six months after the date on which it is received by the Secretary-General of the Council of Europe.

37. In witness whereof the undersigned, duly authorised thereto by their respective governments, have signed the present Convention. Done at Paris, this 14th day of December, 1959, in the English and French languages, both texts being equally authoritative, in a single copy which shall remain deposited in the archives of the Council of Europe. The Secretary-General shall transmit certified copies to each of the signatory and acceding governments.

38. For the Government of the Kingdom of Greece:
At the time of signing the present Convention, I declare that the Greek Government reserves the right not to apply to its own nationals the provisions of Article 3 of the Convention.

Treaty 33 Protocol to the European Convention on the Equivalence of Diplomas Leading to Admission to Universities

39. Protocol to the European Convention on the Equivalence of Diplomas Leading to Admission to Universities
The member states of the Council of Europe signatory hereto, considering the aims of the European Convention on the Equivalence of Diplomas Leading to Admission to Universities, signed at Paris on 11th of December 1953, hereinafter referred to as 'the Convention';
considering that the benefits of the Convention could usefully be extended to holders of diplomas constituting a requisite qualification for admission to universities when such diplomas are awarded by institutions which another contracting party officially

sponsors outside its own territory and whose diplomas it assimilates to those awarded within its territory, have agreed as follows:

40. Article 1

1. Each contracting party shall recognise for the purpose of admission to the universities situated in its territory, when such admission is subject to state control, the equivalence of diplomas awarded by institutions which a contracting party officially sponsors outside its own territory and whose diplomas it assimilates to those awarded within its territory.
2. Admission to any university shall be subject to the availability of places.
3. Each contracting party reserves the right not to apply the provisions of paragraph 1 above to its own nationals.
4. Where admission to universities situated in the territory of a contracting party is outside the control of the state, that contracting party shall transmit the text of this Protocol to the universities concerned and use its best endeavours to obtain the acceptance by the latter of the principles stated in the preceding paragraphs of this Article.

41. Article 2

Each contracting party shall provide the Secretary-General of the Council of Europe with a list of institutions officially sponsored by it outside its territory which award diplomas constituting a requisite qualification for admission to universities situated in its territory.

42. Article 3

For the purpose of this Protocol:
(a) the term 'diploma' shall mean any diploma, certificate or other qualification, in whatever form it may be awarded or recorded, which constitutes a requisite qualification for admission to a university;
(b) the term 'universities' shall mean:
  (i) universities;
  (ii) institutions regarded as being similar in character to universities by the contracting party in whose territory they are situated;
(c) the term 'territory of a contracting party' shall mean the metropolitan territory of that party.

43. Article 4

1. Member states of the Council of Europe who are contracting parties to the Convention may become contracting parties to this Protocol either by:
(a) signature without reservation in respect of ratification or acceptance;
(b) signature with reservation in respect of ratification or acceptance, followed by ratification or acceptance.
2. Any state which has acceded to the Convention may accede to this Protocol.
3. Instruments of ratification, acceptance or accession shall be deposited with the Secretary-General of the Council of Europe.

44. Article 5

1. This Protocol shall enter into force one month after the date on which two member states of the Council of Europe shall have signed it without reservation in respect of ratification or acceptance, or shall have ratified or accepted it in accordance with the provisions of Article 4.
2. In the case of any member state of the Council of Europe who shall subsequently sign the Protocol without reservation in respect of ratification or acceptance, or who shall ratify or accept it, the Protocol shall enter into force one month after the date of such signature or after the date of deposit of the instrument of ratification or acceptance.
3. In the case of any acceding state, the Protocol shall enter into force one month after the date of deposit of the instrument of accession. Such accession shall not, however, become effective until the Protocol shall have entered into force.

45. Article 6

1. This Protocol shall remain in force indefinitely.
2. Any contracting party may, in so far as it is concerned, denounce this Protocol by means of a notification addressed to the Secretary-General of the Council of Europe.
3. Such denunciation shall take effect six months after the date of receipt by the Secretary-General of such notification.

46. Article 7

The Secretary-General of the Council of Europe shall notify the member states of the Council and any state which has acceded to this Protocol of:
(a) any signature without reservation in respect of ratification or acceptance;
(b) any signature with reservation in respect of ratification or acceptance;
(c) the deposit of any instrument of ratification, acceptance or accession;
(d) any date of entry into force of this Protocol, in accordance with Article 5 thereof;
(e) any notification received in pursuance of the provisions of Articles 2 and 6.

47. In witness whereof the undersigned, being duly authorised thereto, have signed this Protocol. Done at Strasbourg this 3rd day of June 1964 in English and French, both texts being equally authoritative, in a single copy which shall remain deposited in the archives of the Council of Europe. The Secretary-General of the Council of Europe shall transmit certified copies to each of the signatory and acceding states.

48. For the Government of the Kingdom of the Netherlands:
"In the case of the Kingdom of the Netherlands, the term 'metropolitan territory' in Article 3(c) of the Protocol shall not retain its original sense but shall be taken to mean 'European territory', in view of the equality in public law between the Netherlands Surinam and the Netherlands Autilles."

49. Declaration of interpretation

At the time of signature of the Protocol to the European Convention on the Equivalence of Diplomas Leading to Admission to Universities, the committee of ministers made the following interpretative clause:
"The Protocol shall also apply to European schools whose certificates fulfil the conditions laid down in Article 1, paragraph (1) of the Protocol."

APPENDIX D

TITLES OF PUBLIC GENERAL ACTS OF PARLIAMENT 1973 - 1975

1973 Statutes

7301. Consolidated Fund Act.

7302. National Theatre and Museum of London Act.

7303. Sea Fish Industry Act.

7304. Atomic Energy Authority (Weapons Group) Act.

7305. Housing (Amendment) Act.

7306. Furnished Lettings (Rent Allowances) Act.

7307. Concorde Aircraft Act.

7308. Coal Industry Act.

7309. Counter-Inflation Act.

7310. Consolidated Fund (No. 2) Act.

7311. Fire Precautions (Loans) Act.

7312. Gaming (Amendment) Act.

7313. Supply of Goods (Implied Terms) Act.

7314. Cost in Criminal Cases Act.

7315. Administration of Justice Act.

7316. Education Act.

7317. Northern Ireland Assembly Act.

7318. Matrimonial Causes Act.

7319. Independent Broadcasting Authority Act.

7320. London Cab Act.

7321. Overseas Pensions Act.

7322. Law Reform (Diligence) (Scotland) Act.

7323. Education (Work Experience) Act.

7324. Employment of Children Act.

7325. Succession (Scotland) Act.

7326. Land Compensation Act.

7327. Bahamas Independence Act.

7328. Rate Rebate Act.

7329. Guardianship Act.

7330. Sea Fisheries (Shellfish) Act.

7331. Dentists (Amendment) Act.

7332. National Health Service Re-organisation Act.

7333. Protection of Wrecks Act.

7334. Ulster Defence Regiment Act.

7335. Employment Agencies Act.

7336. Northern Ireland Constitution Act.

7337. Water Act.

7338. Social Security Act.

7339. Statute Law (Repeals) Act.

7340. Appropriation Act.

7341. Fair Trading Act.

7342. National Insurance and Supplementary Benefit Act.

7343. Hallmarking Act.

7344. Heavy Commercial Vehicles (Control and Regulations) Act.

7345. Domicile and Matrimonial Proceedings Act.

7346. International Cocoa Agreement Act.

7347. Protection of Aircraft Act.

7348. Pakistan Act.

7349. Bangladesh Act.

7350. Employment and Training Act.

7351. Finance Act.

7352. Prescription and Limitation (Scotland) Act.

7353. Northern Ireland (Emergency Provisions) Act.

7354. Nature Conservancy Council Act.

7355. Statute Law Revision (Northern Ireland) Act.

7356. Land Compensation (Scotland) Act.

7357. Badgers Act.

7358. Insurance Companies Amendment Act.

7359. Education (Scotland) Act.

7360. Breeding of Dogs Act.

7361. Pensioners' Payments and National Insurance Act.

7362. Powers of Criminal Courts Act.

7363. Government Trading Funds Act.

7364. Maplin Development Act.

7365. Local Government (Scotland) Act.

7366. Channel Tunnel (Initial Finance) Act.

7367. Fuel and Electricity (Control) Act.

7368. International Sugar Organisations Act.

7369. Northern Ireland Constitution (Amendment) Act.

1974 Statutes

7401. Consolidated Fund Act.

7402. Appropriation Act.

7403. Slaughterhouses Act.

7404. Legal Aid Act.

7405. Horticulture (Special Payments) Act.

7406. Biological Weapons Act.

7407. Local Government Act.

7408. Statutory Corporations (Financial Provisions) Act.

7409. Pensions (Increase) Act.

7410. Representation of the People Act.

7411. Charlwood and Horley Act.

7412. Consolidated Fund (No. 2) Act.

7413. Representation of the People (No. 2) Act.

7414. National Insurance Act.

7415. Consolidated Fund (No. 3) Act.

7416. Independent Broadcasting Authority Act.

7417. Rabies Act.

7418. Contingencies Fund Act.

7419. Lord High Commissioner (Church of Scotland) Act.

7420. Dumping at Sea Act.

7421. Ministers of the Crown Act.

7422. Statute Law (Repeals) Act.

7423. Juries Act.

7424. Prices Act.

7425. Lord Chancellor (Tenure of Office and Discharge of Ecclesiastical Functions) Act.

7426. Solicitors Amendment Act.

7427. Education (Mentally Handicapped Children) (Scotland) Act.

7428. Northern Ireland Act.

7429. Parks Regulation Amendment Act.

7430. Finance Act.

7431. Appropriation (No. 2) Act.

7432. Town and Country Amenities Act.

7433. Northern Ireland (Young Persons) Act.

7434. Pakistan Act.

7435. Carriage of Persons by Road Act.

7436. Mines Working Facilities and Support Act.

7437. Health and Safety at Work etc. Act.

7438. Land Tenure Reform (Scotland) Act.

7439. Consumer Credit Act.

7440. Control of Pollution Act.

7441. Policing of Airports Act.

7442. Independent Broadcasting Authority (No. 2) Act.

7443. Merchant Shipping Act.

7444. Housing Act.

7445. Housing (Scotland) Act.

7446. Friendly Societies Act.

7447. Solicitors Act.

7448. Railways Act.

7449. Insurance Companies Act.

7450. Road Traffic Act.

7451. Rent Act.

7452. Trade Union and Labour Relations Act.

7453. Rehabilitation of Offenders Act.

7454. Pensioners' Payments Act.

7455. National Theatre Act.

7456. Prevention of Terrorism Act.

7457. Consolidated Fund (No. 4) Act.

7458. Social Security Amendment Act.

1975 Statutes

7501. Consolidated Fund Act.

7502. Education Act.

7503. Arbitration Act.

7504. Biological Standards Act.

7505. General Rate Act.

7506. Housing Rents and Subsidies Act.

7507. Finance Act.

7508. Offshore Petroleum Development (Scotland) Act.

7509. Supply Powers Act.

7510. Satute Law (Repeals) Act.

7511. Social Security Benefits Act.

7512. Consolidated Fund (No. 2) Act.

7513. Unsolicited Goods and Services (Amendment) Act.

7514. Social Security Act.

7515. Social Security (Northern Ireland) Act.

7516. Industrial Injuries and Diseases (Old Cases) Act.

7517. Industrial Injuries and Diseases (Northern Ireland Old Cases) Act.

7518. Social Security (Consequential Provisions) Act.

7519. Export Guarantees Amendment Act.

7520. District Courts (Scotland) Act.

7521. Criminal Procedure (Scotland) Act.

7522. Oil Taxation Act.

7523. Reservoirs Act.

7524. House of Commons Disqualification Act.

7525. Northern Ireland Assembly Disqualification Act.

7526. Ministers of the Crown Act.

7527. Ministerial and Other Salaries Act.

7528. Housing Rents and Subsidies (Scotland) Act.

7529. Mental Health (Amendment) Act.

7530. Local Government (Scotland) Act.

7531. Malta Republic Act.

7532. Prices Act.

7533. Referendum Act.

7534. Evidence (Proceedings in Other Jurisdictions) Act.

7535. Farriers (Registration) Act.

7536. Air Travel Reserve Fund Act.

7537. Nursing Homes Act.

7538. Export Guarantees Act.

7539. Hearing Aid Council (Extension) Act.

7540. Diseases of Animals Act.

7541. Industrial and Provident Societies Act.

7542. New Towns Act.

7543. British Leyland Act.

7544. Appropriation Act.

7545. Finance (No. 2) Act.

7546. International Road Haulage Permits Act.

7547. Litigants in Person (Costs and Expenses) Act.

7548. Conservation of Wild Creatures and Wild Plants Act.

7549. Mobile Homes Act.

7550. Guard Dogs Act.

7551. Salmon and Freshwater Fisheries Act.

7552. Safety of Sports Grounds Act.

7553. Public Service Vehicles (Arrest of Offenders) Act.

7554. Limitation Act.

7555. Statutory Corporations (Financial Provisions) Act.

7556. Coal Industry Act.

7557. Remuneration, Charges and Grants Act.

7558. Lotteries Act.

7559. Criminal Jurisdiction Act.

7560. Social Security Pensions Act.

7561. Child Benefit Act.

7562. Northern Ireland (Emergency Provisions) (Amendment) Act.

7563. Inheritance (Provision for Family and Dependants) Act.

7564. Iron and Steel Act.

7565. Sex Discrimination Act.

7566. Recess Elections Act.

7567. Housing Finance (Special Provisions) Act.

7568. Industry Act.

7569. Scottish Development Agency Act.

7570. Welsh Development Agency Act.

7571. Employment Protection Act.

7572. Children Act.

7573. Cinematograph Films Act.

7574. Petroleum and Submarine Pipe-lines Act.

7575. Policyholders Protection Act.

7576. Local Land Charges Act.

7577. Community Land Act.

7578. Airports Authority Act.

7579. Consolidated Fund (No. 3) Act.

7580. OECD Support Fund Act.

7581. Moneylenders (Crown Agents) Act.

7582. Civil List Act.

7583. Northern Ireland (Loans) Act.