Unsupervised Categorization of Word Meanings Using Statistical and Neural Network Methods

Christopher Cedric Huckle

Ph.D.

The University of Edinburgh

1996

To my family, with thanks for their encouragement and support over many years

ACKNOWLEDGEMENTS

I am grateful to my supervisors, Dr Richard Shillcock and Dr Hamish Macleod, for their help whilst I was carrying out the work described in this thesis. I should also like to thank Dr Nick Chater, now at the University of Oxford, for giving me the opportunity of pursuing a Ph.D. in the first place, and the Carnegie Trust for the Universities of Scotland for their financial support. Finally, I wish to acknowledge my debt to Julian Smith for his patient assistance during the early stages of my research.

ABSTRACT

This thesis investigates the extent to which the statistical structure of natural language can be used to enable a conceptual structure for word meanings to be developed without external supervision.

Few of the words that human beings use can be assigned rigourous definitions, yet we all have an intuitive understanding of the relationships between the meanings of these words. Psychology has yet to provide a complete account of how this knowledge is obtained. The present work seeks to extend recent statistical approaches to syntactic development to consider the problem of developing a categorization for word meanings, using techniques which have recently been popular in the fields of Computational Linguistics and Neural Computation.

A statistical technique is introduced for representing the contexts in which words occur. Each word is represented by a 'statistical context vector', and the vectors are subjected to hierarchical cluster analysis to produce a structure in which words which have similar contexts are placed closer together than those which do not. Analyses of this type are carried out on a 10,000,000 word corpus taken from the Wall Street Journal, using a variety of different parameters, and the appropriateness of the resulting structures is assessed using Roget's Thesaurus as a benchmark.

A still more attractive approach is one which deals with polysemy, and which develops its representations for word meanings continuously from the outset, with no need for a separate stage of statistical analysis.

To take these considerations into account, an unsupervised neural network is presented, in which different senses of a word token are allowed to be assigned to different output clusters as the contexts of their occurrence dictate. After initial testing using Elman's (1988) artificial corpus, the network's performance is assessed on the 10,000,000 word corpus by comparing the ways in which different word tokens are distributed over the output units.

Further analyses are carried out in which a crude measure of this distribution is used to assess Jones' (1985) 'Ease of Predication' measure. The distribution measure is found to account for a significant amount of the variance in Ease of Predication. Word frequency is also found to play a significant role, and word frequency effects are reconsidered in the light of this. The psychological implications of the results obtained from the network are discussed.

It is concluded that there is a great deal of information inherent in the structure of language which could potentially play an important part in developing a conceptual structure for word meanings. Whilst extralinguistic information is undoubtedly likely to be of importance as well, it is striking that the use of very simple statistical measures can permit the development of such rich structures.

TABLE OF CONTENTS

1. INTRODUCTION

1.1 Overview

The work described in this thesis was carried out in the Department of Psychology and the Centre for Cognitive Science at the University of Edinburgh. It provides links between two relatively unconnected disciplines; Psychology and Computational Linguistics, and presents various existence proofs of the usefulness of statistical information in learning the relationships between word meanings.

We shall first introduce and discuss the psychological problem of accounting for our ability to categorize words on the basis of meaning. We shall then investigate this empirically using statistical techniques which have most often been used by researchers in Computational Linguistics seeking improved methods for natural language processing (as opposed to gaining insight into aspects of language acquisition in human beings).

This thesis is, indeed, primarily an empirical endeavour, attempting to investigate in an objective manner aspects of a problem which has often occupied the minds of those working in Psychology. The extensive computer programming work involved in the analyses presented later was carried out in the C programming language, using various UNIX facilities available at the University of Edinburgh.

1.2 Outline

We shall begin our investigation of the role of statistical information in categorizing word meanings by considering the psychological background to the problem in Chapter 2. We shall consider, for example, some of the ways in which psychologists have attempted to account for the ability of human beings to categorize the words they know, and we shall confront questions such as that of whether supervised learning is likely to be useful in performing such a categorization. In Chapter 3, we shall review the approaches which have been taken, mainly within Computational Linguistics, in trying to make use of the statistical structure contained within natural

language. We shall also consider the utility of such approaches in addressing the psychological issues outlined in Chapter 2.

In Chapter 4, a number of statistical analyses are carried out and the results presented in the form of tables of nearest neighbours and as dendrograms. A more objective evaluation of these analyses will then be introduced and implemented in Chapter 5, where we shall see that semantic analyses present problems for evaluation which are more acute than for syntactic analyses which use similar methods.

In Chapter 6, we shall reassess the whole approach taken in Chapter 4 and identify some particular problems associated with it. In particular, we shall be concerned with the need to allow more than a single representation for each of the words considered. In Chapter 7, these considerations are used to motivate the development of an unsupervised neural network capable of categorizing natural language in an on-line fashion. To assess whether the behaviour of the network does actually reflect these considerations, its performance will first be assessed when applied to a simple artificial corpus containing a restricted set of syntactic categories. Having established that the network is indeed able to perform well with this corpus, in Chapter 8 it is applied to the problem of categorizing words from a large natural language corpus. We shall also consider some of the psychological aspects of its performance and compare these to an existing psychological metric. In Chapter 9, we shall conclude by considering the possible ways in which future work might build on the work presented here. In particular, we shall consider ways in which the neural network introduced in this thesis might yield further psychological insights.

1.3 Terminology

Whilst we shall attempt to avoid the use of much technological terminology in this thesis, there is one pair of terms which will frequently recur, and which are worth explaining at the outset. These are the terms 'target word' and 'context word'.

We shall be using the term 'target word' at various points in the thesis to denote the word whose representation we are concerned with. The term 'context word', on the

other hand, refers to a word used to represent a target word; in other words, it is a word in terms of which the target word is being represented. Usually we shall use more than one context word in the representation of a target word. If we chose, to take a simplified example, to represent the word 'tiger' in terms of the words 'stripe', 'India', and 'ferocious', then 'tiger' would be our target word, and the context words would be 'stripe', 'India', and 'ferocious'. Of course, the specific details surrounding the use of the various representations we shall encounter will be explained in detail at the appropriate juncture.

2. *Natural Language Categorization and Psychology*

2.1 The Need for Categorization

Human beings are exposed continuously to large amounts of information, received in various forms by the different sensory modalities. However, it is clear that we have the ability to classify and to organize this information to enable it to be used more efficiently than would be possible if we were to attempt to use it in its raw form (see, for example, Barlow (1989) and Redlich (1993) for reviews of approaches to achieving this).

In doing so, we may often find it useful to use some kind of *summary* in order to concentrate on those features of an entity that happen to be important for a particular task or in a particular context, and to de-emphasize those which are less important. Thus, when informing a salesman that we wish to purchase the 'red' carpet, it is probably not relevant that the carpet also contains small patches of purple and has yellow borders in order to know which carpet is being referred to. The fact that it is predominantly red will probably serve to *distinguish* it from one that is, say, predominantly blue. Of course, we would have to change our summary here if the range of available carpets included more than one predominantly red style; the 'red' carpet could not then be so easily identified.

Using a similar sort of technique, we may also ignore some of the ways in which entities differ in order to make use of the fact that they are also *similar* in certain important ways. Thus, we might speak of the 'red' carpets when distinguishing them from others, even though the predominantly red carpets in question might not be identical. The characteristic of being predominantly red is, in this case, sufficient to allow us to distinguish these carpets from others which are predominantly blue.

In making sense of the world around us, and in sharing our knowledge about the world with others, we frequently make use of these sorts of generalisations, which

exploit the statistical redundancy in the world around us, for simplifying the data with which we are confronted. Pinker (1994) notes one aspect of the usefulness of this:

"Lumping objects into categories — giving them a category label in mentalese — allows one, when viewing an entity, to infer some of the properties one cannot directly observe, using the properties one *can* observe. If Flopsy has long furry ears, he is a "rabbit"; if he is a rabbit, he might scurry into a burrow and quickly make more rabbits (p155)."

The formal mathematical analysis of this domain is provided by information theory (see Shannon and Weaver (1963), for example, for a useful introduction). In this thesis, we shall be looking at the usefulness of the sorts of techniques we have described above in enabling us to classify *words* on the basis of similarities and dissimilarities in the contexts in which they are used. As Ritter and Kohonen (1989) observe, such classifications are an important and universal element of natural language:

"The most general concepts or abstractions that are needed to interpret the empirical world are called *categories*; such basic reduced elements and forms of thinking and communication can also be encountered in all languages, primitive as well as more developed (p241)."

The assumption will be made here that, if context is a useful source of information about semantics, then, generally speaking, words which are similar with respect to the contexts in which they occur ought to have greater similarity in their meaning than words which are not so similar in terms of context. We shall, of course, be making this notion much more formal as we proceed, but we will nonetheless be making use of the sorts of summary techniques outlined here to deal with the information that is contained within language data.

2.2 The Psychological Literature on Concepts and Word Meanings

The approach taken in this thesis towards considering the utility of context in developing conceptual structure for word meanings will be largely empirical. It will focus mainly on one issue: the extent to which the information provided by the statistical structure within the language is useful in allowing knowledge of the relationships between word meanings to be developed. However, it should be noted that psychologists and psycholinguists have, using various different methodologies, produced an extensive literature concerned with the characteristics of conceptual

structure. Whilst it is not the intention here to give an exaustive consideration of this research (see Van Mechelen, Hampton, Michalski and Theuns (1993) for a comprehensive review), we shall now outline the sorts of approaches that have been taken, and consider some of the questions which have been addressed.

Garnham (1990) provides a useful review of psychological approaches to explaining the way in which word meanings are represented and organized by human beings. Garnham has noted that early psycholinguistic attempts to account for word meanings made use of *features* to represent each word. Chomsky (1965), for example, proposed that sets of bivalent features called semantic markers, such as 'MALE/FEMALE' and 'HUMAN/NON-HUMAN', could be used in this way. To account for lexical ambiguity (the phenomenon whereby a given string of letters can refer to more than one concept), Chomsky suggested that each sense of a word had its own accompanying set of semantic markers, but that 'distinguishers' were also present to enable the senses to be differentiated.

An alternative to the approach of regarding word meanings as represented in terms of sets of features was proposed by Collins and Quillian (1969), who introduced to Psychology the notion of semantic networks. Here, the word meanings are represented as nodes interconnected by a hierarchical network of links representing various relations between the meanings, such as 'is dangerous' and 'can move around'. The word's meaning is characterised by the position of its node relative to the rest of the network. Gallant (1991) has noted, however, that semantic networks are not likely to be an ideal solution to the problem of creating a representation for word meanings; various different types of links would be required in such a network to allow the rich variety of possible context effects to be transmitted between nodes, but this would in turn require a vastly complicated network. (For an overview of more recent connectionist implementations of semantic networks, see, for example, Lange (1992)). Feature-based semantic representations are nonetheless still often assumed to be appropriate. For example, in a recent review of connectionist approaches to modelling human reading abilities, Plaut, McClelland, Seidenberg, and Patterson (1996) note that:

"We imagine that the semantic representations for words are relatively sparse, meaning that each word activates very few of the possible semantic features and each semantic feature participates in the meanings of a very small percentage of words (p105)."

Both feature theories and semantic networks seem to suggest, however, that words can be defined in terms of a set of necessary and sufficient conditions. In an extensive review of the area, Medin and Smith (1984) have described those accounts in the literature on concepts which make such suggestions as the 'classical' view. They point out a number of criticisms of this view. Firstly, it has, in practice, turned out not to be possible to list the defining properties for concepts despite many years of attempts to do so; relatively few categories ('bachelor' is often used as an example here) have clear cut membership criteria - for instance, is a fake lion a member of the 'lion' category? Secondly, human subjects' judgements about the boundaries between the categories into which words fall appear to be more flexible than the classical view might suggest; for example. Thirdly, some members of such categories are judged to be more typical of those categories than others. Fourthly, the frequency of occurrence of the properties of a particular word appears to be correlated with its perceived typicality as a member of the class into which it falls; since 'has a beak' and 'can fly' are properties frequently associated with members of the 'bird' category, a word which possesses these properties is likely to be judged as a typical member of that category. Fifthly, it has been observed that *non*-necessary attributes may be used by subjects in placing particular words into categories. Finally, the classical view presupposes that the defining properties of a word must be included in those of a 'superordinate' word, and that the superordinate word will additionally be defined in terms of some further properties. Thus, 'bird' should include all the properties of 'sparrow', plus some additional properties. It would follow from this that a word should be judged as more similar to an immediate superordinate than to a distant one. However, Medin and Smith note that this has not always been observed to be the case in empirical studies.

In the 1970's, a number of psychologists came to regard the requirement of a list of necessary and sufficient properties for a word as an inappropriate assumption in defining its meaning, leading Armstrong, Gleitman, and Gleitman (1983) to assert that:

"generally speaking, it is widely agreed today in philosophy, linguistics, and psychology, that the definitional program for everyday lexical categories has been defeated - at least in its pristine form ... (p268)".

As a consequence, psychologists started to embrace the theory of prototypes (see, for example, Rosch (1973)). The idea here is that our concepts of word meanings are represented as being clustered around the representation of a *prototypical* member of the category that the word denotes. The prototype (examples of which may not exist in reality) contains more of the properties of the category than marginal members, and items which contain more properties will accordingly be judged as better exemplars of the category than those which contain fewer. Such an approach is related to Wittgenstein's (1953) observation that, since the meanings of words cannot be defined in terms of necessary and sufficient attributes, the most appropriate way of thinking about their meaning is in terms of overlapping sets of similarities or as sharing 'family resemblances'. By thinking in these terms, the finding that some words are judged to be 'better' exemplars of a category than others seems easier to accommodate than with the classical definitional view of concepts[1].

More recent experimental approaches to the psychological study of concepts and the properties which they possess include the work of McRae (1992), who investigated the role of the intercorrelations between the properties in accounting for the representation of real-world concepts. In McRae's investigation, subjects were asked to produce lists of properties for various exemplars of 'natural kind' categories (such as mammals and vegetables) and for various exemplars of 'artifact' categories (such as kitchen items and vehicles). Using these data, concepts in which the properties were strongly intercorrelated were compared in a decision task with those in which the properties were only weakly intercorrelated. In the decision task, subjects were shown a concept, followed by a property, and were asked to respond as quickly as possible whether the property was 'reasonably true' of the concept which preceded it. The results revealed that decision latencies here were significantly faster if there were strong intercorrelations between property shown and the properties of the concept

[1] Armstrong, Gleitman, and Gleitman (1983) have shown, however, that graded membership judgements can also be obtained for those words which presumably *do* have clear definitions, such as 'female'; 'chairwoman' and 'cowgirl' were judged by subjects to be much poorer exemplars here than 'housewife' and 'mother'. This was felt to raise some interpretative difficulties for studies which use paradigms requiring judgements of this kind.

than if there were relatively weak intercorrelations. This was felt to be the influence of a pattern completion process in which, when the representation of a property is accessed, properties correlated with it also become activated. Further investigations suggested that there was a distinction between the assessment of similarity between the natural kind concepts, which was felt to involve the intercorrelation of properties, and the assessment of similarity between artifact concepts, which involved independent properties.

McRae, de Sa and Seidenberg (1993) followed up these empirical studies indicating that human beings encode correlations between properties of real-word entities. McRae et al. produced a model of conceptual memory which was implemented using a Hopfield neural network (Hopfield (1982)). Since this type of network uses a correlational learning rule, it was seen as a natural means with which to encode correlations between the properties of the entities it encounters. McRae et al. described their approach in these terms:

"Our goal was to use an explicit computational model to investigate a theory in which encoded knowledge of the correlational structure of semantic space plays a critical role in computing concepts from words (p729)."

The input units of the network permitted a distributed representation of the words' spelling. The output representations were based on conceptual norms produced by subjects, with one output unit for each of 646 of these norms. Following training, the network learned to produce the appropriate conceptual representation for 80 of the 84 words with which it was presented, and further investigations of the network's performance supported the findings of McRae (1992). In particular, it was noted that the intercorrelational density of a concept's properties did significantly influence the activation strength of those properties.

2.3 An Operational Definition for Word 'Meaning'

It is intuitively clear, and experiments such as those of McRae et al. (1993) support this, that human beings have some kind of organized conceptual structure for word meanings. We are able to identify words that are similar in meaning, such as 'doleful' and 'forlorn', and are able to contrast them with or distinguish them from words that

are quite different in meaning, such as 'portmanteau' and 'ecstatic'. As Lakoff (1989) makes clear, our ability to classify words in this way represents an important aspect of our cognitive abilities:

"... whether they are used in nonlinguistic tasks or not, linguistic categories *are* categories - and they are part of our overall cognitive apparatus. Whether one wants to dignify them with the term "conceptual" or not, linguistic categories are categories within our cognitive system and a study of *all* categories within our cognitive system will have to include them (p88)."

What is perhaps not always quite so obvious is that comparisons and contrasts of the kind just indicated are, in the majority of cases, *all* we can do when attempting to identify the meaning of words and when trying to convey information about word meanings to others. For most words, rigourous definitions do not exist; as noted above, this has been a major difficulty for the classical definitional view of representing word meanings. Whilst a physicist may be able to refer to 'force' *precisely* as 'the rate of change of momentum' (assuming the words in this definition have equally exact definitions themselves!), most words used outside such restricted domains as physics cannot be defined so easily. In fact, most dictionary definitions boil down to the sorts of comparisons and contrasts we have just encountered, and are thus somewhat recursive in character. As Garnham (1990) points out,

"dictionaries define one word in terms of others, and semantic memory must represent relations among word meanings too. However, dictionaries and semantic memories have different purposes and make different uses of the interconnections between words. Dictionaries are consulted to ascertain the meaning of unknown words. Their entries assume that the meanings of other words are known, and use known words to explain the meaning of new words ... Someone who wants to know the meaning of *tamarack* wants to know what kinds of things tamaracks are. A dictionary gives this information by assuming that its users are familiar with the things mentioned in the definition. (p114)".

The following two examples, taken from the 1989 edition of Chambers English Dictionary, may help to make this clear. Firstly, the definition for 'ensnare':

"to catch in a snare: to entrap: to entangle."

Secondly, the definition for 'entity':

"being, existence: something with objective reality: an abstraction or archetypal conception".

Such definitions as these require one first to have acquired the meanings for the words contained within them. By relating the two words concerned to the words in the definition, an attempt is then made to convey their meaning. Yet it is also clear that

these definitions would not in themselves be a sufficient basis for capturing the understanding of their meaning as it is possessed by a native speaker of English.

The variety of words used by human beings which can be used to refer to the same thing indicates the closeness in the perceived meaning of groups of these words (though it does not, of course, follow that the meanings are perceived as identical), and emphasizes the complexity of our conceptual structure for word meanings. This fact emerges as a practical difficulty when human beings interact with computer systems, and has been described as the 'vocabulary problem' by Furnas, Landauer, Gomez and Dumais (1987); there is a very wide variation between the words used by individuals to refer to information they wish to access, which can often result in ineffective search techniques.

We have noted that it is usually not possible, in practice, to describe the meaning of most words in precise terms. The question might then reasonably be asked, however, as to what the meaning of a word would look like if it *could* be given (since we do, after all, appear to be in possession of such information). We shall be talking about 'meanings' a great deal, and it is clearly important to state what we intend by the notion of 'meaning'.

To deal with this, we shall not attempt to supply any kind of very formal definition of what the meaning of a word is. Instead, we shall adopt the operational approach of regarding the meaning of a word as the knowledge which would be sufficient to determine when that word could be used appropriately in the language in question to communicate about a particular concept (whether concrete or abstract) and when it could not. This corresponds closely to describing the *sense* of the word, the traditional semantic term used by Frege to refer to the set of objects which a particular word denotes; however, knowledge of this kind, as we have seen, cannot usually be articulated for the words that we typically use. This is so *even though* we ourselves as native speakers of a particular language are capable of making the distinction between correct and incorrect usage and thus act *as if* we are in possession of the required knowledge. As Wetter and Nüse (1992) put it:

"the variety of semantics and pragmatics of utterances is faintly felt but not systematically understood (p436)".

Our operational definition for meaning here is similar to Wittgenstein's (1953) assertion that the meaning of a word is in its use. Wittgenstein captured the idea of a word's meaning in the following terms:

"For a *large* class of cases — though not for all — in which we employ the word "meaning" it can be defined thus: the meaning of a word is its use in the language (Wittgenstein, 1953; §43)".

The notion that our understanding of the meaning of a word is reliant upon knowledge of its use is, of course, a major reason for the difficulty in communicating the meaning of 'meaning'; it is a word like many others, for which we may be able to discern the correct and incorrect usages, whilst being unable to supply a statement to describe these rigourously. As Garnham (1990) puts it:

"the ability to understand is not the same as being able to explicate the concept of meaning ... (p96)".

Smadja (1989) has also noted this aspect of the meaning of words, and has suggested that 'co-occurrence knowledge' represents a distinct type of lexical knowledge, which

"represents the extent to which an item is specified by its environment independently of syntactic or semantic reasons (p163)".

It is this type of knowledge, Smadja asserts, which allows us to assess the well-formedness of utterances, and to observe the proper usage of words. Given its important role in our knowledge of word meanings, Smadja stresses that such information should form part of computational dictionaries.

One of the central themes for this thesis, as we shall see below, is to explore some of the ways in which we may come to be in possession of this incommunicable understanding of the meanings of the words we use.

2.4 Developing a Structure for Word Meanings

A major concern of this thesis is the *development* of the conceptual structure, rather than just a characterisation of its nature. This conceptual structure for word meanings must, to a large extent, be learned. One major reason for believing this comes from examination of the concepts held by children, which are often different from those used by adults. Mervis (1989) has described various ways in which the disparity can

occur. The child's categories can be broader than those of an adult, as, for example, when the word 'kitty' is used to refer to tigers and lions as well as cats, and when 'ball' is used to refer to rounded objects such as beads and piggy banks as well as items adults would describe as balls. The child can, alternatively, be more specific in the use of categories than an adult typically would be, as when not including beanbags in the category labelled by the word 'chair'. A third possibility is that the child's category may overlap with some of the adult's category, but not include all items the adult would include; the child's understanding of the category for 'car' might, for instance, include lorries but not include beach buggies.

Mervis notes that this difference between the adult and child lexicons provides a challenge for theories of lexical development:

"An adequate theory should be able to explain why such differences consistently occur and how these differences eventually disappear; that is, how the extension of a word for a child comes to correspond to the extension of that word for an adult. Ideally, a theory should be able to predict when these differences will appear and disappear, rather than accounting for them post hoc. (Mervis, 1989; p201)."

The debate surrounding the issue of language acquisition has, of course, often occupied the minds of psychologists. It is not an aim of this thesis to attempt to resolve this issue or to provide a complete theory of lexical development. Instead, it is the intention to explore empirically an approach which concentrates on learning to classify words on the basis of meaning from scratch using intralinguistic sources of information only. However, we shall briefly consider the points arising from this debate.

The central question in the debate surrounding language acquisition has been that of the extent to which the language experience of language learners contributes to their language development. One of the most notable contributions to the debate has arisen from the conflict which occurred in the 1950's between the behaviourist approach of Skinner, and the nativist approach of the linguist Chomsky. Skinner (1957) had claimed that children learn language through the processes of operant conditioning, and discounted any necessity for the child to have prior knowledge about language.

To take two notable examples concerning language acquisition from Skinner's account, he described the acquisition of a 'tact' as occurring through situations in

which, in the presence of a stimulus, a child will emit a verbal response to receive reinforcement. Thus, when seeing a doll, the child may utter the word "doll" in order to gain some sort of reinforcement. On the other hand, acquisition of what Skinner described as a 'mand' would occur when a child is carrying out some activity and hears the word "No!". In such a situation, the child must stop the activity in order to receive positive reinforcement or to avoid some kind of aversive stimulation. According to Skinner, the next time the activity is carried out, the child is likely to receive reinforcement for uttering the word "No!" himself.

Addressing the problem of understanding how it is that words come to refer to entities in the world, and thus come to have meanings, Skinner felt that learning processes like these would provide the key:

> "... the speaker will emit a response of a given form in the presence of a stimulus having specified properties under certain broad conditions of deprivation or aversive stimulation. *So far as the speaker is concerned*, this *is* the relation of reference or meaning. There would be little point in using this formula to redefine concepts such as sign, signal, or symbol or a relation such as reference, or entities communicated in a speech episode such as ideas, meanings, or information. These traditional terms carry many irrelevant connotations ... (Skinner, 1957; p115)".

One of Chomsky's main challenges to Skinner's account was that it could not satisfactorily explain one of the major aspects of language acquisition: the development of the ability to produce and to comprehend utterances which have not previously been heard. Chomsky (1965) proposed that the acquisition of a language must involve the application of a set of rules, and that some innate knowledge about language in general must be supposed. The innate knowledge was seen as necessary because Chomsky believed that the linguistic experience of the language learner would not be sufficient to permit the language to be acquired correctly, even if that experience consisted entirely of correct examples of the language being learned. In addition to this, Chomsky argued that children are simply not given the systematic reinforcement which Skinner's account would require them to have; in fact, he claimed that the language heard by children is often 'degenerate', being affected by the linguistic performance of the speakers, and would *not* provide a good model of the correct usage of the language.

As noted above, the intention here is to explore the extent to which an empirical approach which works *only* by using the structure within the language data itself can be successful in permitting a categorization for word meanings to be acquired. If it should be that this information is useful here, it would suggest that an efficient system for learning language should make use of it. At the same time, we are not suggesting that human language learners restrict themselves to the use of this information alone, or that the behaviourist account of language acquisition is to be fully endorsed. It is worth bearing in mind the comments of Redington, Chater, and Finch (1995) here:

"After the development of generative grammar distributional linguistics was justly criticized on a number of grounds ... including connections with dubious doctrines such as behaviorism and positivism, lack of formal rigor, a failure to properly deal with syntax, and an over-restrictive definition of linguistics, which ruled out semantics, and any psychological aspects of language. We suspect that the bad name of distributional linguistics has led many researchers to discount the possibility that distributional information of any sort can have any bearing on language and language acquisition (p6)"

2.5 The Role of Context

To explore the usefulness of the information source provided by language data, we shall later be considering the similarities and differences between words on the basis of the contexts in which they occur. The study of context effects has been considered from various perspectives in Psychology and Psycholinguistics. Tanenhaus and Lucas (1987), for example, have reviewed some of the evidence for the influence of context on lexical processing. In particular, they have assessed the reasonableness of maintaining that the various subsystems (such as syntax and semantics) involved in lexical processing are functionally autonomous, acting as independent modules. If they are, context effects ought not to arise because the processing of each module would not pay attention to the processing being carried out in other modules. Tanenhaus and Lucas distinguish between pre-lexical context effects, which influence word recognition, and post-lexical context effects, which are concerned with the selection and integration of lexical representations which have already been activated.

As Tanenhaus and Lucas (1987) have noted, psychologists have often used *priming* techniques to examine post-lexical context effects. Such techniques make use of a paradigm in which a subject is shown a prime word followed by a target word. The

subject must then decide whether or not the target word is a real English word. Of particular interest to this thesis is Tanenhaus and Lucas' consideration of the effects of semantic context here. Lexical decision latencies have been found to be shorter when the prime word is related to the target word. For example, when the prime word 'bread' is followed by the target word 'butter', the lexical decision speed is faster than when the prime word is 'nurse'. Psychologists have felt it an important issue to establish the reason for this kind of effect. Tanenhaus and Lucas point out that one possible explanation for the priming effect is that the two words involved share some semantic 'features', and that the features activated when the meaning of the first word is accessed may feed back to activate the lexical representation for the second word. If this account is correct, the systems involved cannot be modular in nature because of the fact that information is being fed back from one level of the process to another[2].

On the other hand, the semantic priming effect can be explained in a way which allows the view of the lexical processing systems involved as modular ones to be preserved; the idea here, due to Fodor (1983), is that whatever things are encountered frequently in the real world will give rise to correspondingly strong connections between relevant nodes in the mental lexicon. A problem with this, however, is that it is a more convincing account of *associative* priming (which obtains between words which frequently co-occur in a language) than it is of semantic priming (which often obtains between words which do *not* frequently co-occur in this way). This difficulty lends some credence to the notion that it is association with a particular *concept* that may make words similar, rather than merely the association of occurrence between each other.

There is an extensive literature within Psychology and Psycholinguistics on the subject of priming, and we shall not attempt a thorough exploration of it in this thesis. However, it should be noted that priming studies represent a major branch of empirical research aimed at exploring the effects of linguistic context on word recognition and access to word meanings. Zwitserlood (1989) has provided an influential discussion here, having carried out cross-modal priming experiments which

[2] Tanenhaus and Lucas (1987) define a modular system as one in which feedback is not allowed between different levels of a system.

suggest that, whilst the sentential semantic context preceding a prime word is used to choose between contextually appropriate and contextually inappropriate target words, the effect of this does not come about until after the prime itself has begun to be heard (but before the sensory information provided by hearing it is sufficient to disambiguate between the target words), by which stage various lexical candidates will have been activated. Thus, the process of disambiguation would appear to be one of selecting between a 'cohort' of various candidate words on the basis of sentential context and sensory information, as opposed to one of using context to preselect a single candidate. For a review of the cohort model of spoken word recognition, see, for example, Marslen-Wilson (1987). More recently, Moss and Marslen-Wilson (1993) have reported findings which contradict such an interpretation, which they describe as the 'exhaustive' account of access to word meanings. Moss and Marslen-Wilson found, again using the cross-modal priming paradigm, that the accessed *properties* of prime words varied according to the preceding biasing sentential context, suggesting that such context does affect initial lexical access. This view is also supported by Tabossi and Zardon (1993), who present experimental findings suggesting that context can enable selective access of the appropriate meaning of ambiguous words

Spivey-Knowlton, Tanenhaus, Eberhard and Sedivy (1995) have recently used an experimental paradigm involving an eyetracker to provide evidence that visual context does influence subjects' recognition of verbal instructions given to them. When the speech signal was still ambiguous as to the object being referred to, subjects often fixated their gaze on an object in the visual environment whose name began with the same sequence of segments as that which they had so far heard. Thus, in the instruction "pick up the penny", subjects might fixate initially on a pencil in front of them, since 'penny' and 'pencil' are members of the same cohort of lexical candidates when only the first syllable has been heard. This finding does suggest a close coordination between spoken language and the extralinguistic environment, as does related work by Sedivy, Tanenhaus, Eberhard, Spivey-Knowlton, and Carlson (1995).

As we have already noted, it is not our intention to look at priming studies in great detail here. In this thesis, we shall restrict ourselves to an examination of the

intralinguistic structure which may allow conceptual structure for word meanings to be *developed*, and which could go some way to providing an explanation of the mechanics which bring about priming effects. In Chapter 3 we shall describe the work of Bullinaria and Huckle (1996) in putting this to the test by using statistical vectors derived from real natural language corpora as inputs to a neural network model of semantic and associative priming.

The influence of context has been examined using paradigms other than that of priming. Harris (1992), for example, has provided a comprehensive review of empirical investigations into the influence of language itself on the development of the understanding of various word meanings. The acquisition of words was monitored for four children from the age of 6 months to 2 years. Among the earliest words acquired, 'context-bound' words were identified, which were used by the child in the same context on each occasion of use. In such cases, the context was often characterized by the child carrying out the same action each time the word was used. For example, one of the children, the word 'choo-choo' was used only when the child was pushing a toy train along the floor. In addition to such words, however, more contextually flexible words were also identified which were used to refer to a wide range of objects sharing some element in common. For example, the word 'shoes' was used by some of the children to refer to different types of shoe right from the beginning of the use of this word. Similarly, 'more' was used in various ways, such as when taking more bricks from a toy box, when about to take another drink from a cup, and when holding out an empty bowl at meal times.

What is of particular interest are Harris' investigations of the relationship between the use of a word by the mother and the initial use of that word by the child. It was found that the relationship was quite a close one in terms of the contexts in which the words were used by the mother and the child; the child's initial use often closely resembled that of the mother, and in only 3 out of 40 initial word uses considered was there no identifiable relationship between the mother's and the child's usages. Furthermore, in 33 of these cases, the child's initial use of a word was related to the most frequent usage of that word by the mother during the preceding month. It was found, however,

that the subsequent use of words was considerably less closely related to the mother's usage than initially. Harris concluded that:

"the onset of vocabulary production is thus firmly rooted in the children's experience of adult speech. And, more particularly, it is rooted in the experience of hearing particular words frequently being used in particular contexts (p91)".

Harris also noted that when the children employed words which were initially context-bound in their usage, these words gradually showed an increasing flexibility in terms of the contexts in which they were used. This was particularly so for 'nominal' words, which were used to refer to objects or classes of objects.

This empirical work suggests, then, that there is evidence for a close correspondence between the content and frequency of items in the linguistic input (provided in this case primarily by the mother's speech), and their use by children learning the language. Harris notes that there are two possible explanations for this:

"One is that the child's observation of maternal word use had a direct influence on production, that is, the child modelled his/her own initial use of a word on maternal word use. The other possibility is that maternal word use had an indirect influence via comprehension. That is to say, the child's experience of maternal word use determined the contexts in which a word was understood before production, and the pattern of comprehension then determined the initial pattern of production (p90)".

The second possibility here is one which will be of particular relevance for the approach taken in this thesis.

2.6 Unsupervised Learning

The empirical work outlined above supports the intuition that only a limited amount of the development of an understanding for word meanings is achieved by supervised learning. This is a form of learning in which the learner approaches the solution to a problem by being informed of the solution each time he or she makes an attempt to solve that problem (see, for example, Rumelhart, Hinton, and Williams (1986)). By reducing the 'distance' between the attempted, incorrect solution, and the correct solution, the learner will gradually produce responses which become closer and closer to the desired correct one. Experimental evidence does suggest that for common words, the repeated corrections of an adult can serve to teach a child the referents of various common objects. Mervis (1989), for example, confirms that some

categorization is achieved by children before they are able to comprehend language. Pinker (1994), furthermore, notes that children exhibit some inbuilt tendencies to consider only certain types of categorization when confronted with novel objects. These tendencies rule out many inappropriate ways of carrying out the categorization, easing the learning task with which the children are confronted. It is worth pausing to consider, however, whether a supervised approach to learning could account for our knowledge of the meanings of *all* the words we use. This would require repeated corrections to be made by adults for each of the 20,000 or so words the child will acquire, and thus appears to be utterly implausible as a general explanation for acquisition of word meanings.

The difficulties presented by supervised learning become particularly clear when we consider *abstract* words. For these words, such as 'similar' and 'justice', concrete referents do not exist, and so it is even less easy to attempt to define them than concrete words such as 'puddle' and 'scarf'. As Miller (1963) states:

> "Abstract nouns cause much trouble because we learn to use them without learning the full sequence of symbols they are intended to replace. Then we begin to disagree with others on how to replace them and perhaps decide eventually that no acceptable definition exists (p111)".

Nonetheless, it remains true that we have strong intuitions about the usage of abstract words and can readily categorize them on the basis of similarity in meaning. However, we typically cannot supply rigourous definitions for them, and would have great difficulty in imparting their meaning to others by error correction. Whilst it might be possible to do so using reinforcement learning, in which supervision simply takes the form of information as to whether the learner's response is correct or incorrect, time constraints would seem to rule this out for the majority of the abstract words acquired by human beings. It is worth noting that accounting for the representation of the meaning of abstract words has often been ignored by the psychological literature on concepts, which has tended to concentrate on the case of commonplace, concrete words. (See, however, Roitblat and von Fersen (1992) for a review of comparative psychological approaches here which does consider abstract concepts). As Rice (1990) makes clear, descriptions of children's acquisition of word meanings have also tended to neglect the case of abstract words, and have assumed the importance of supervision:

"Current experimental studies of word learning are relatively constrained in terms of the input conditions, the kinds of word meanings explored, and the implicit model of word knowledge. Focused adult input is presumed ... almost all of our expoerimental literature focuses on object and attribute words, most often drawn from carefully proscribed semantic fields (p191)".

The alternative approach, suggested both by intuition and by the work of Harris (1992) is to use unsupervised learning, in which the desired solution to a problem is not provided explicitly. Instead, it is the statistical structure of the input data which provides the required information. In particular, we shall be concerned in this thesis with the statistical relationships between words, making the assumption that the acoustic signal provided by speech has already been segmented. For a review of approaches to dealing with this problem of segmentation, see, for example, Jusczyk (1993).

In a problem requiring data to be structured or classified, the unsupervised approach would achieve this by detecting statistical regularities which render some of the items in the data more similar to each other than to others. The example given earlier of categorizing carpets on the basis of their predominant colour works along the same lines. No explicit information is given about the desired classification to be made (i.e. 'red' carpets versus 'blue' ones). The fact that two carpets are predominantly red provides a statistical basis for grouping them together. Using information of this kind to classify language potentially provides an alternative to the view that our knowledge of word meanings must be part of our genetic endowment, as Ritter and Kohonen (1989) make clear:

"At the time when the genetic predisposition of language elements was suggested, there was no mechanism known that would have explained the origin of abstractions in neural information processing other than evolution. It was not until "neural network" modeling reached the present level when researchers began to discover *internal representations* of abstract properties of signals ... such findings indicate that the internal representations of categories may be derivable from the mutual relations and roles of the primary signal or data elements themselves ... (p242)."

In order for unsupervised learning to be successful in building up the sort of categorization for word meanings discussed earlier, there are three important assumptions which must be made (and investigated):

1. The assumption that the words in question do share some kind of statistical regularities. Whilst unsupervised learning does not require the explicit provision of the

correct solution to the problem, it does require the existence of some form of statistical structure within the data being encountered. If there is no statistical structure here, and entities in the data are related to one another on a purely random basis, then no reliable categorization of the data will be possible.

2. The assumption that these regularities can be *detected* by the system doing the learning. If this is not so, because, for example, the system can only pick up less sophisticated regularities than are going to be required, then the categorization will again fail.

3. The assumption that the use of these regularities will permit the particular categorization we are seeking. We are assuming here that by making use of statistical regularities in the data, and thereby ending up with a categorization of language, we will end up with a categorization which looks like the intuitive one we are familiar with. Whilst we hope that this might turn out to be the case, there is no guarantee that it will be.

We have already conceded that supervision probably plays a role in the learning of word meanings, and it may well be that unsupervised learning alone can come nowhere near the sophistication required to achieve what human beings have achieved. However, this would nonetheless be an informative finding, consistent with one of the central aims of this thesis, which is to investigate *the extent to which* unsupervised methods can permit a categorization of word meanings to be achieved. Since it is not *necessarily* the case that they can, we must rely on some form of empirical investigation.

2.7 Conclusions

In this chapter, we have seen that accounting for the way in which word meanings are acquired and represented is a problem which psychologists have addressed in various different ways. We have seen that any discussion of the notion of 'meaning' presents difficulties, and a working definition for this concept has therefore been proposed. With a view to investigating the acquisition of word meanings, a task at which young

children rapidly make progress, we have also noted that, in general, supervised learning appears to be a less promising approach than unsupervised learning.

In order to develop these broad conclusions, we need to find some method for carrying out unsupervised learning with linguistic input, in order that the statistical structure within the language can be used as a source of information about words and their contexts. This will then permit an exploration of the usefulness of such information in learning about the relationships between the words encountered. Fortunately, although Psychology has not, as yet, made a great deal of use of them, promising empirical methods have been developed in recent years within the field of Computational Linguistics. In Chapter 3, we shall explore some of these and examine how they may be of use in addressing the problems which have been discussed in this chapter.

3. STATISTICAL METHODS IN COMPUTATIONAL LINGUISTICS

3.1 Statistical Methods in Computational Linguistics

The use of distributional methods in natural language research has recently seen a revival within the field of Computational Linguistics. Such approaches had been used within Linguistics in the 1950's; Harris (1954), for example, discussed distributional context in terms which are similar to those used in much more recent investigations, such as those which will be presented in Chapter 4:

> "The distribution of an element will be understood as the sum of all its environments. An environment of an element A is an existing array of its co-occurrents, i.e. the other elements, each in a particular position, with which A occurs to yield an utterance. A's co-occurrents in a particular position are called its selection for that position (p146)."

Harris' approach, however, was concerned chiefly as a means for *describing* the structure of language, and did not direct much of his attention to the practical usefulness of such structure in developing knowledge of word meanings and their interrelations. Similarly, Fries (1952) presented a description of the English language in terms of aspects of its distributional structure. This was based on an analysis of about 250,000 words of English speech recorded in the United States, and was regarded as a scientific alternative to the traditional methods used in sentence analysis during the preceding decades.

There has been considerable recent interest amongst researchers interested in natural language in the use of statistical methods for grouping words (e.g. Sinclair (1991); Charniak (1993)). Such methods have been particularly attractive to those working within the field of Computational Linguistics. There are a number of reasons for this:

1. It is hoped that such approaches will generate more impressive results than traditional natural language processing approaches which need access to real-world knowledge. The provision of such knowledge has turned out to be very problematic for the field of artificial intelligence, and methods which do not require it are therefore tempting.

2. Large machine-readable corpora are now readily available for carrying out the statistical analyses, thereby giving more promise than the much smaller scale analyses carried out by, for example, Fries (1952). Such analyses require to be large-scale in order for the statistical measures used to be reliable. This is particularly important where low frequency items are concerned.

3. Computers which are fast enough to carry out the required analyses in reasonable time are also now available.

The methods adopted within Computational Linguistics are also of particular interest in approaching the psychological problem of developing a conceptual structure for word meanings. Among the reasons for this are that such methods:

1. Potentially provide a means for developing conceptual structure without supervision.

2. Do not require *a priori* distinctions to be drawn between, say, concrete and abstract words; all words can be represented using the same sort of approach.

We have seen in Chapter 2 that there is a question to be answered about the extent to which unsupervised statistical methods can permit a conceptual structure for word meanings to be developed. Much of the statistical work carried out recently within Computational Linguistics is of relevance in providing the empirical means with which to confront this question. To illustrate this, we shall now examine some of the main contributions to this body of work in detail.

In advocating the use of statistical methods, Gallant (1991) argued that representing words in terms of vectors was an attractive approach. Rather than the vector components being probabilities (as has become conventional), however, he suggested that they might assume only a very limited set of values (such as +2 (strongly associated), +1, 0, -1 and -2 (strongly negatively associated)) to indicate the degree to which the word represented by the component was associated with the target word. Gallant estimated that compilation of 200-dimensional vectors for 2000 words would take around 80 person-days to complete if performed manually. Happily, it has since

transpired that vectors of probabilities (rather than only the 5 levels of association) can be created automatically in a small fraction of this time.

Brown, Della Pietra, deSouza, Lai, and Mercer (1992) demonstrated the use of information theoretic procedures to produce semantically coherent groups of words. In addition to this, they investigated the phenomenon of 'sticky pairs' of words which occur in English. These are pairs of words which are more likely to occur adjacent to one another than they are to occur independently of one another. By calculating the mutual information between pairs of words in text taken from the Canadian parliament, Brown et al were able to show that the stickiest pairs did often correspond to familiar constructions in the English language, such as 'Humpty Dumpty', 'Tse Tung', and 'ammonium nitrate'. This approach was then made more general by grouping together sets of words that occurred together more often than would be expected from the statistics of their occurrence in the text, and whose mutual information was thus higher than would be predicted on the basis of chance. Semantically coherent groups of words were found to result, again demonstrating the potential utility of statistical techniques in producing interesting categorizations of natural language.

Schütze and Pedersen (1993) investigated the use of vector representations for words in performing categorizations of language data. This work was conducted from the perspective of enabling those involved in constructing dictionaries to improve the classifications made. By 'improve' here, it appears that Schütze and Pedersen meant a closer approximation to the usages which are encountered in the linguistic performance of human beings, which are of particular interest to this thesis:

> "With the information that can be extracted from large amounts of text, one can hope to discover usages that have eluded the lexicographer's eyes, to make dictionaries more representative of actual language use and to update them more rapidly and more accurately than is possible today (p1)."

Schütze and Pedersen drew a distinction between sets of words which are 'syntagmatic associates' and those which are 'paradigmatic parallels'. They used the former term to refer to words which are frequently neighbours of one another (such as 'he' and 'wrote'), and the latter to refer to words which frequently have similar neighbours to the left or the right (such as 'reduce' and 'cut'). It is worth noting at

this point that these two types of relationships between words correspond closely to the psychological distinction between the relationship which gives rise to associative priming and the relationship which gives rise to semantic priming respectively. Schütze and Pedersen also note that it is paradigmatic parallels which need to be distinguished in dictionary definitions.

Schütze and Pedersen used a corpus of 17 million words to record collocations between the 5,000 most frequent words in the corpus. This resulted in each of the 5,000 target words being represented as a 5,000 dimensional vector whose components contained co-occurrence counts with each of the 5,000 context words. Schütze and Pedersen proposed the use of a singular value decomposition technique was to reduce the dimensionality of the vector space, although how many dimensions were ultimately used is not clear from this particular report. This is desirable from the point of view of reducing the required storage space and also for enabling some undesirable idiosyncracies of the corpus to be deemphasized. Similarity between words could be assessed by measuring the cosine of the angle between their respective vectors. The vector representations for words were also subjected to a clustering procedure. Investigation of syntagmatic associates and paradigmatic parallels revealed that subtle usages of the words concerned could be revealed using these techniques. These were seen to be encouraging examples of the manner in which the techniques used could be of benefit in the construction of dictionaries. The authors also noted that many combinations of the parameters involved remained unexplored.

Schütze (1993a) used similar techniques to examine the possibility of extracting the parts of speech present in language data. This was also the objective of Finch and Chater (1992a, 1992b) working from a more psychological perspective, and their work is discussed further below. Schütze's motivation for investigating the issue included the fact that lists of part-of-speech labels may not be available for less frequent words in English, or for words in languages other than English, or for words occurring within a particular genre of text. It would therefore be desirable on occasion to be able to generate these labels from scratch.

The corpus used by Schütze (1993a) was taken from the New York Times, and the most frequent 5,000 words were used as target words. Vector representations were derived for each of these words. The vectors contained co-occurrence counts for context words occurring at each of four positions relative to the target word (last word, next word, last-but-one word, next-but-one word). The dimensionality of these vectors was reduced to 15 dimensions using a singular value decomposition procedure. It was discovered that the nearest neighbours to the words considered were very often of the same syntactic class.

Schütze (1993a) also used this approach using vectors of co-occurrences with word *classes* rather than individual words. This was done to permit a larger proportion of the vocabulary of such a large corpus to be dealt with; the singular value decomposition technique would have been prohibitively time consuming if carried out for very large numbers of words in the corpus. The word classes to be used were obtained by clustering the original vectors (containing *word* co-occurrences) into 500 clusters. These clusters were used as 500 features with which to represent 22,771 target words from the corpus. The 500 dimensional space was then reduced to 10 dimensions using the singular value decomposition procedure.

Again, a random sample of the target words showed that their nearest neighbours were very often members of the same syntactic category, indicating that the contextual information latent within the vectors was able to reflect the syntactic similarities between target words. It should be noted, however, that semantic similarities were also evident in the neighbouring words listed by Schütze, hinting that semantic classification might be possible using a similar procedure.

3.2 Bringing Psychological and Computational Linguistic Methods Together

Little work has been carried out to date in bringing together the two areas of interest discussed in this chapter and the preceding one. Chapter 2 outlined the psychological problem of categorizing natural language, and introduced the psychological literature on the subject, whilst the present chapter has so far been concerned to demonstrate

the utility of various statistical approaches used within Computational Linguistics in investigating the categorization of language, particularly syntactic categorization. One of the main concerns of this thesis, however, is to explore some of the ways in which the latter methods may be applied to the former problem.

A limited amount of work has been carried out by some researchers in bringing together the psychological and the computational aspects of the problem of natural language categorization. Finch and Chater (1992a, 1992b) and Lund, Burgess, and Atchley (1995), among others, have done so from the perspective of Psychology or Cognitive Science. These authors have to an extent been interested in the language acquisition debate in Psychology. However, their work has, with the exception of Lund, Burgess, and Atchley (1995) mainly been concerned with *syntactic* categorization.

Rather than deriving their investigative techniques from those employed in Computational Linguistics, however, Finch and Chater were initially interested in research within Cognitive Science which used supervised neural networks for syntactic categorization. This interest appears mainly to have been due to the work of Elman (1988), which is discussed in detail in Chapter 7.

The investigation of unsupervised semantic categorization by computational means has not, with the exception of some work undertaken by Finch and Chater (1995), received recent attention from those adopting a psychological perspective to the problem; it has, as we have seen above, received rather more attention from those working within more computational disciplines.

This thesis will make a contribution here. As we noted earlier in Chapter 2, the intention is to focus on the single, intralinguistic, source of information provided by the language data alone in order to try to obtain useful insights regarding its influence on the conceptual structure.

As Harris (1992) has noted, such an approach can often be criticized unfairly:

"... any suggestion that language development might be influenced by linguistic input is sometimes mistakenly seen as returning to an empiricist position that was demolished long ago (p3)".

By way of defence against this sort of charge, it should be pointed out that it is our intention to examine the utility of this single source of information in allowing a categorization of word meanings to be developed, without attempting to claim that it is the only candidate source of information. If it transpires that it is indeed useful, we may perhaps be able to reach a measured conclusion such as that of Hughes (1994), who, following a demonstration that distributional statistics can be useful in the acquisition of *syntax*, notes that this

"... demonstrates that some structure can be extracted on the basis of distributional redundancy. This is not meant to constitute proof of empiricist acquisition of any sort. It merely counters the claim that no language structure can be acquired from sparse and variable data (p135)".

Dealing with the same issue, Redington, Chater and Finch (1995) point out that:

"simple distributional methods are sometimes associated with a general empiricist *tabula rasa* approach to language learning, which has been widely criticized ... However, this is not germane in the present context, since distributional methods are not proposed as a general solution to the problem of language learning, but rather as a possible source of information about syntactic structure (p3)".

It is in this kind of vein that the present work is undertaken. The statistical information contained within natural language is a potential source of information about the development of word meanings which has not yet been thoroughly researched, despite the fact that it is a conspicuous source of potentially useful information. It is also, given the power of contemporary computing resources, a very accessible one. There may very well be other aspects of such development which need to be understood. However, for the present, we shall concern ourselves only with the data provided by the language itself.

As we have already noted, the areas of distributional statistical analysis of language, and that of answering psychological questions about language acquisition have not, until very recently, been integrated. As Redington, Chater and Finch (1995) note of the distributional work carried out prior to the Chomskian revolution:

"Distributional linguists were interested in the discovery of language structure from corpora, purely from the point of view of providing a rigorous methodology for field linguistics; they did not consider that this approach might have any relationship to language acquisition in children (p6)"

Wolff (1976, 1988) advocated an approach towards the study of the acquisition of syntax and semantics which is very much in the spirit of the work carried out in this thesis, using information about the frequencies of word co-occurrences to capture the contexts in which words occur and thereby to categorize language into words, or syntactic and semantic structures. However, this work was restricted through being carried out on a much smaller scale than more recent corpus-based approaches have permitted.

Of the recent work which has been carried out in bringing these two strands together, much has been due to the work of Finch and Chater (e.g. Finch and Chater (1991, 1992a, 1992b)), involving acquisition of syntactic structure. This work represents a significant contribution to the research that has been conducted in the area, and we shall therefore now consider their work in detail.

Finch and Chater (1991) introduced their interest in the use of simple statistical techniques as providing a potentially important means for learning the structure of natural language, in contrast to some others who would claim that such methods are unable to capture the very rich structure in language. Finch and Chater (1991) were mainly concerned with the use of neural network methods in exploiting statistics inherent in the language data, and this work is discussed more fully in Chapter 7.

Finch and Chater (1992a, 1992b), however, applied some statistical methods not implemented in the form of a neural network. The aim of their work was to investigate the psychological problem of 'bootstrapping' syntactic categories. This problem arises from the fact that, in the absence of prior knowledge about a particular domain (such as syntax), it is extremely difficult to learn about the structure of that domain. In such a situation, we would not have an appropriate set of categories into which to place the phenomena being observed, nor would we know anything about the rules defined over those categories. We would be faced with the necessity of learning both the categories and the rules from scratch. The 'bootstrapping' problem is, as Finch and Chater (1992a, 1992b) point out, one in which

"... the specification of a set of rules presupposes a set of categories, but the validity of a set of categories can only be assessed in the light of the utility of the set of rules that they support (Finch and Chater (1992b), p230)".

In their particular case, the bootstrapping problem is one of learning the set of syntactic categories and the grammatical rules which are defined over them.

Finch and Chater (1992a, 1992b) did not have the benefit of 'prior knowledge' in the form of, for example, a text corpus containing words tagged with the appropriate syntactic categories; instead, they approached the problem with an untagged corpus. They then sought to *derive* syntactic categories using simple statistical measures applied to the corpus. It was hoped that, using their statistical approach, a similarity measure could be reached which would reflect useful underlying syntactic categories. Once the categories had emerged in this fashion, the next stage of such an approach would be to tag the corpus using them, and then attempt to determine rules defined over them:

"Thus a hierarchy of categories and rules can be derived by iterating this process. This method also promises to allow the revision of initial categorisation decisions, based on impoverished assumptions concerning the set of rules, in the light of the rules derived ... (Finch and Chater 1992b, p321)".

To reach the similarity measure they hoped to obtain, Finch and Chater (1992a, 1992b) recorded statistical information about the *context* of use of the words they were studying. They defined the context for each word simply as the two words preceding it and the two words following it. This approach was motivated by the existence of the 'replacement test', which is a standard justification in Linguistics for the syntactic categories conventionally assigned to words in natural language. This test asserts that a word or a phrase which can be *replaced* by a word or a phrase of a known category is itself a word or a phrase belonging to that category. Finch and Chater's approach, then, was presented as a means of operationalising this test, allowing it to be applied empirically. They described this operationalisation as the 'statistical replacement test':

"Has the word or phrase been observed to occur in a corpus in similar contexts to another word or phrase? If so, then these should be given similar linguistic categories (Finch and Chater (1992b), p321)."

Finch and Chater (1992a, 1992b) restricted their attention to the 1000 most frequent words in the corpus, with context information about the 150 most frequent words in

the corpus being recorded. For each of the 1000 words being considered, the context information was recorded in the form of 4 vectors each of 150 dimensions. Each of these vectors corresponded to one of the context positions relative to the word being considered (preceding word, following word, last word but one, next word but one), and each of the dimensions within it corresponded to one of the 150 context words.

The distance metric used by Finch and Chater (1992a, 1992b) for establishing the similarity between these vectors was the Spearman Rank Correlation Coefficient, a metric which was found to give rise to categories which accord well with conventional syntactic categories, and which has the important and desirable property of being insensitive to the absolute frequency of the words being represented by the vectors. The inter-vector distances calculated using the Spearman metric were then used as the input for a hierarchical cluster analysis procedure, in which words having similar vectors would be placed closer together in the hierarchy than those having relatively more dissimilar vectors.

Finch and Chater (1992a, 1992b) used a corpus of 40,000,000 words taken from the Usenet for their analyses. The dendrogram containing the 1000 words investigated was found to contain nodes resembling a conventional taxonomy of syntactic categories; the nodes in the tree were found to correspond closely to these categories, such as determiners, prepositions, verbs, adjectives and so on. In addition to this syntactic structure, some semantic structure was also revealed in the dendrogram produced. The semantic structure was revealed within clusters of words sharing the same syntactic category. Thus, within the part of the dendrogram containing a large number of adverbs, semantically related words were found to be clustered together; examples of this phonemenon included 'finally' and 'eventually', 'thus' and 'therefore', 'maybe' and 'perhaps', and 'never' and 'always'. Similarly, a group of nouns was identified all of which related to computing: 'software', 'data', 'text', 'modem', 'cable', 'font', and 'utility', for example. Other clusterings of nouns revealed groups of words sharing some semantic aspects, such as words related to nations of the world, numbers, and world leaders.

Redington, Chater and Finch (1993) applied these methods to 4,300,000 words taken from the CHILDES corpus (MacWhinney and Snow, 1985), which consists of children's speech and child-directed speech. This was also found to produce interesting results, with the resulting dendrogram having classified the 1000 most frequent words in the corpus (which were represented in terms of 150 context words) into appropriate clusters. It was found that 13 clusters accounted for more than 90% of these words, with the labels for the clusters being deemed appropriate for about 95% of the words in each cluster. Semantic groupings were once again in evidence, and to illustrate this, Redington et al. (1993) provided examples of food related words here ('water', 'milk', 'food', 'cake', 'egg', and so on).

The analysis was also performed on child speech and adult speech separately, with more appropriate classifications being observed for the latter. This was explained by the noisier data provided by the child speech.

Redington et al. (1993) also investigated the ability of the approach to cope with novel words presented in context, which should, in principle, be 'recognized' as being similar in nature to some of the words already encountered. The vectors for the new words were constructed as before, with four vectors of 150 dimensions used to represent the context. The single occurrence of the new word meant that these vectors were extremely sparse, with a maximum of one non-zero entry in each of the four context vectors. These vectors were compared to the context vectors for each of the words previously considered, using the dot product between them as a distance metric. The word was 'recognized' as belonging to the category of the previously encountered word for which the dot product was largest. It was found that nouns and verbs could be dealt with to a surprising extent, although novel words which were not members of these categories were not distinguished nearly as well. Redington et al.'s approach in dealing with novel words here shows some similarity to the neural network implementation for handling ambiguity which is discussed in detail in Chapter 7.

This work was further developed by Redington, Chater and Finch (1995), underlining their emphasis on the potential importance of simple distributional statistics in acquiring knowledge of the syntactic categories into which words in natural language fall. They have stressed once again, however, that the use of such a method need not be the only source of information here, specifying four other possible sources. These are: relating the linguistic input to the communicative context in which it occurs, considering phonological cues, analysing prosody, and using innate knowledge of syntactic categories.

Redington et al.'s first analysis involved the use of an artificial stochastic context free grammar comprising 1,000,000 words taken from a vocabulary of 111 items. All words were used as target and context items, and distances between the vectors resulting from the analysis were calculated using the Cityblock metric. As with Finch and Chater (1992a, 1992b), information was recorded about the statistical behaviour of context words in each of four positions relative to the target word (previous word, previous word but one, next word, next word but one). A hierarchical cluster analysis procedure was then carried out, producing a dendrogram which revealed a perfect correspondence between its own branches and the syntactic categories of the original grammar. Two items which were ambiguous as to syntactic category were found to appear in the dendrogram between the entries for the two possible branches in which they might be placed.

The amount of information conveyed by the dendrogram about the syntactic categories specified in the original grammar was then quantified. This was achieved by calculating the mutual information between the categories revealed by cutting the dendrogram at each of a number of levels, and the categories in the original classification. Redington et al.'s analysis here revealed that at all levels of cutting the dendrogram, apart from the extreme cases where all items fall into a single category and where each item has its own category, the divisions in the dendrogram conveyed more information about the original grammatical categories than would be expected by chance.

Similar analyses were carried out on a part of the CHILDES corpus (MacWhinney and Snow, 1985), consisting of more than 2,500,000 words. For the purposes of the analysis, the most frequent 1,511 words were used as target words and the most frequent 150 words were used as context words. Evaluation of the resulting classification was not quite so straightforward as with the corpus involving the use of the artificial grammar because in this case the syntactic designation of each item in the corpus could not so easily be determined. To deal with this difficulty, the evaluation was facilitated by referring to the most common syntactic category assigned to each of the words by the Collins Cobuild lexical database[3]. As with the artificial grammar, it was found that cutting the resulting dendrogram at any interesting level conveyed more information about the syntactic categories assigned to the words from the database than would be expected on the basis of chance. Interestingly from the point of view of the concerns of this thesis, semantic information was also revealed in the dendrogram; Redington et al. reported groupings, for example, of food-related nouns ('chocolate', 'gum', 'toast', 'corn', 'carrot', and so on), and of nouns related to parts of the body ('hands', 'feet', 'legs', 'ears', 'teeth', and so on).

Redington et al. do suggest that distributional approaches may be a reasonable way to acquire information about word meanings. They note the importance of the work of Gleitman (1994), who argues that syntactic information may be important in acquiring verb meanings, and that since syntactic information appears to be obtainable by distributional means, it may follow that semantics can be reached in this way too.

A further recent approach which is of interest is that of Zavrel and Veenstra (1995). In attempting to provide some empirical evidence that bootstrapping of syntactic categories may be possible from the structure of language itself, and that a 'prewired representational system for syntactic categories' would not therefore be required, Zavrel and Veenstra used vector representations for the most frequent 5000 words in a 3,000,000 word corpus taken from the Wall Street Journal.

[3]Redington et al.'s method was not able to assign any word to more than one syntactic category (although syntactic ambiguity is commonly encountered in natural language). This considerable disadvantage is discussed further in Chapter 6.

Depending on the analysis being conducted by Zavrel and Veenstra, each word was represented with a vector containing information about either the most frequent 250 words or the most frequent 1000 words in the corpus. This information took the form of a frequency count of the number of times the word occurred in a neighbouring position to the word being represented, normalized by the frequency of the word being represented, or in some analyses by the total frequency of the contextual word itself. As with Finch and Chater (1992a, 1992b) the four neighbouring positions were considered. Zavrel and Veenstra, whose method of evaluation will be discussed in detail in Chapter 5, demonstrated that using their straightforward means for representing the statistical behaviour of words in relation to their neighbours, syntactic clusters could indeed be identified. That is, words which would conventionally be assigned the same syntactic category would tend to lie close to one another in the space defined over the vectors representing them.

The work we have considered so far in this section has been concerned with syntactic categorization of language. The interest of this thesis is, however, largely that of semantic categorization. The work of Lund, Burgess and Atchley (1995) is of particular relevance here. These authors aimed to establish whether or not a correlation exists between the results of semantic priming experiments in Psychology and the distances measured between words when represented by statistical vectors. This approach, then, differs from many of those described hitherto in that it is concerned with using statistical methods to investigate psychological issues relating to semantics. Lund et al. described their approach, which they regarded as a *model* of various semantic effects encountered in the cognitive and neuropsychological literature, as the Hyperspace Analogue to Language (HAL). Since Lund et al.'s paper is so similar in approach to the material to be covered in this thesis, we shall consider it in detail.

Lund et al. analysed an English text corpus of 160,000,000 words taken from the Usenet, making use of a moving window 10 words in length. This length of window was justified on the basis that it would not be so small as to miss constructs that span several words (such as long noun phrases), and yet would not be so large as to introduce too many extraneous co-occurrences. Within the 10 words spanned by the

moving window, co-occurrences were recorded for each of the 70,000 most frequent words in the corpus. Each time a pair of these words was found to co-occur within the moving window, their co-occurrence value was incremented by one, weighted by the distance between the words. Thus, the co-occurrence count of a word pair separated by a gap of 9 words was weighted by a 'co-occurrence strength' of 1, while the same pair occurring adjacent to one another was given a weight of 10.

Recording these co-occurrences resulted in a 70,000 x 70,000 matrix of counts. Since each row contained information about the extent to which each of the 70,000 words preceded the word represented by that row, and each column contained information about the extent to which each of the 70,000 words followed the word represented by that column, each word's full representation would consist of both the row and the column for that word, resulting in a 140,000 element vector. It is unclear from Lund et al.'s paper, however, how the row and the column for each word were combined.

Lund et al. reported that examination of the variance across the columns in the new vector revealed that the variance dropped sharply across the first hundred elements, and was very low by the two hundredth element. As a result of this, the 139,800 columns containing the lowest variance (that is, all but the first 200 columns for each word) were discarded.

As noted above, the manner in which the 140,000 element vector for each word was constructed from the two original 70,000 element vectors is unclear from the paper's description of the procedure. Lund (*personal communication*) has, however, described the manner in which this was done. The two 70,000 element vectors (from the row and the column of the 70,000 x 70,000 element matrix) were concatenated, with the first 70,000 elements corresponding to the elements of the original row vector, and the second 70,000 elements corresponding to the elements of the original column vector. This resulted in a 140,000 element vector for each of the 70,000 words being considered. In each of the experiments carried out, a subset of these 140,000 element vectors was used, corresponding to whichever words were being investigated in that particular experiment.

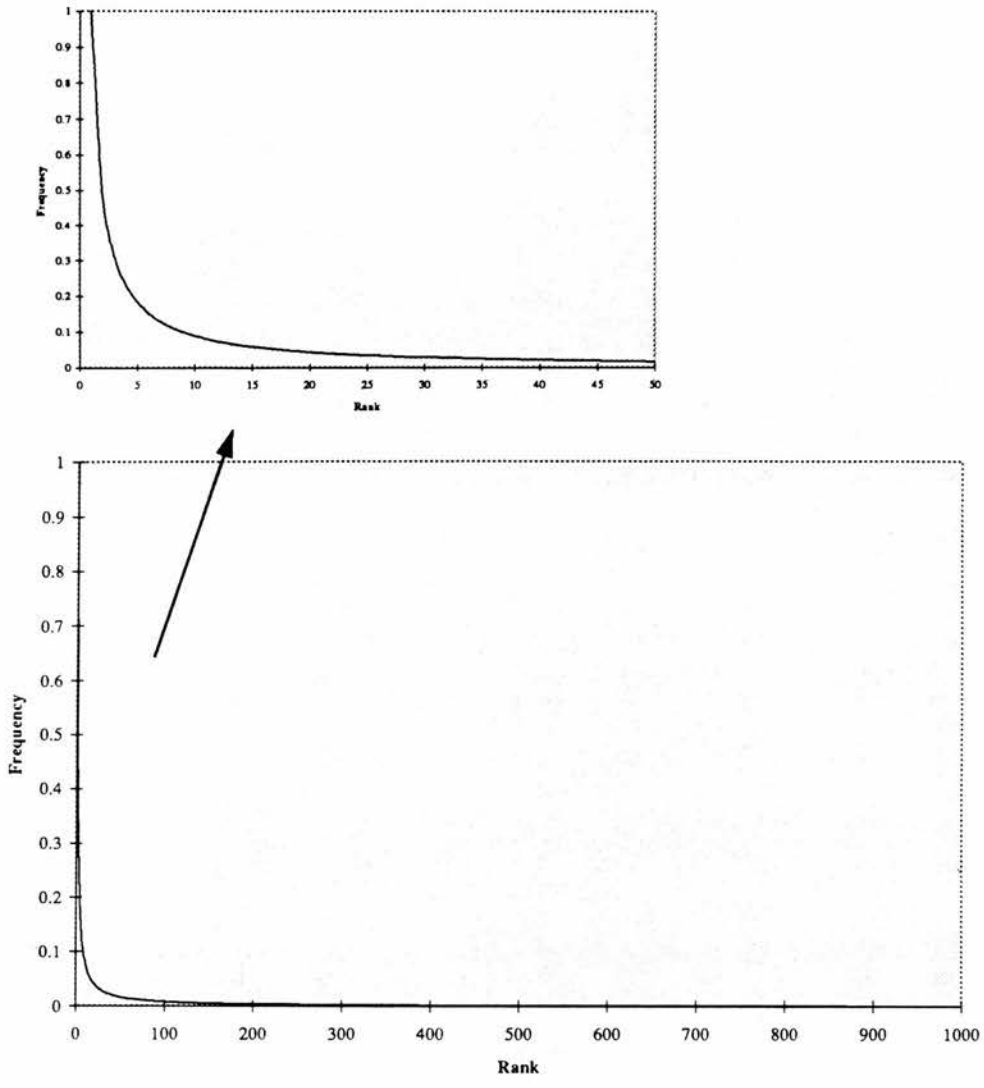
These vectors were next arranged as a matrix in which each row was 140,000 elements in length, and in which there were as many rows as there were words being investigated. Lund then explains that the 140,000 columns of this matrix were arranged in decreasing order of variance, with those columns having variance lower than some criterion being discarded. It turned out that only 200 columns were retained, while 139,800 were discarded.

It is surprising that Lund et al. did not foresee this before they embarked on collecting statistics. Collecting co-occurrence information for the most frequent 70,000 words in the corpus is highly unusual; most researchers would gather this information for the most frequent 5,000 or fewer words. For the most frequent 70,000 words to give useful results, a huge corpus would be required. This reasoning can be supported by Zipf's law, Mandelbrot's refinement of which states that (see Hughes (1994), Mandelbrot (1983)):

$$f_n \propto n^{-1.05} \tag{3.1}$$

where n is the rank of the nth most frequent item in the text, and f_n is the frequency of that item. Thus, in approximate terms, the product of a word's frequency and its rank is a constant. This gives a characteristic curve, indicating that the majority of the items in the text each contribute relatively little to its total number of words. Such a curve is plotted in figure 3.1 for 1000 items and (inset) 50 items.

Figure 3.1: Theoretical Relationship Between Word Frequency and Rank



This is a phenomenon which corpus language research has to recognize, because many words in a text will have low frequencies, and may not therefore provide reliable statistics in a given corpus. For this reason, researchers have tended to include only the most frequent few hundred or thousand items in their analyses.

The use of the most frequent 70,000 words in the corpus by Lund et al. is surprising because it would be expected that the vast majority of these words would have very low frequencies, introducing this risk of unreliability. Furthermore, since the curve above has a gradient close to zero for many of these words, many of their frequencies will also be very similar to one another. Thus there are two problems inherent in this

particular approach. The first is that many of the words will be too infrequent to provide reliable statistics. The second is that the similarity between the frequency counts for many of them will mean that their inclusion in the analysis will be relatively uninformative; as Redington et al. (1995) confirm,

"... bigram frequencies, like word frequencies, follow Zipf's law (p10)".

Since Lund et al. combined their vectors in the manner outlined earlier, the variance across many of the columns (those corresponding to all but the most frequent items[4]) would be expected to be low for these reasons. It is not surprising, therefore, that they felt it appropriate to discard 139,800 columns on the basis that they were uninformative; as the authors note,

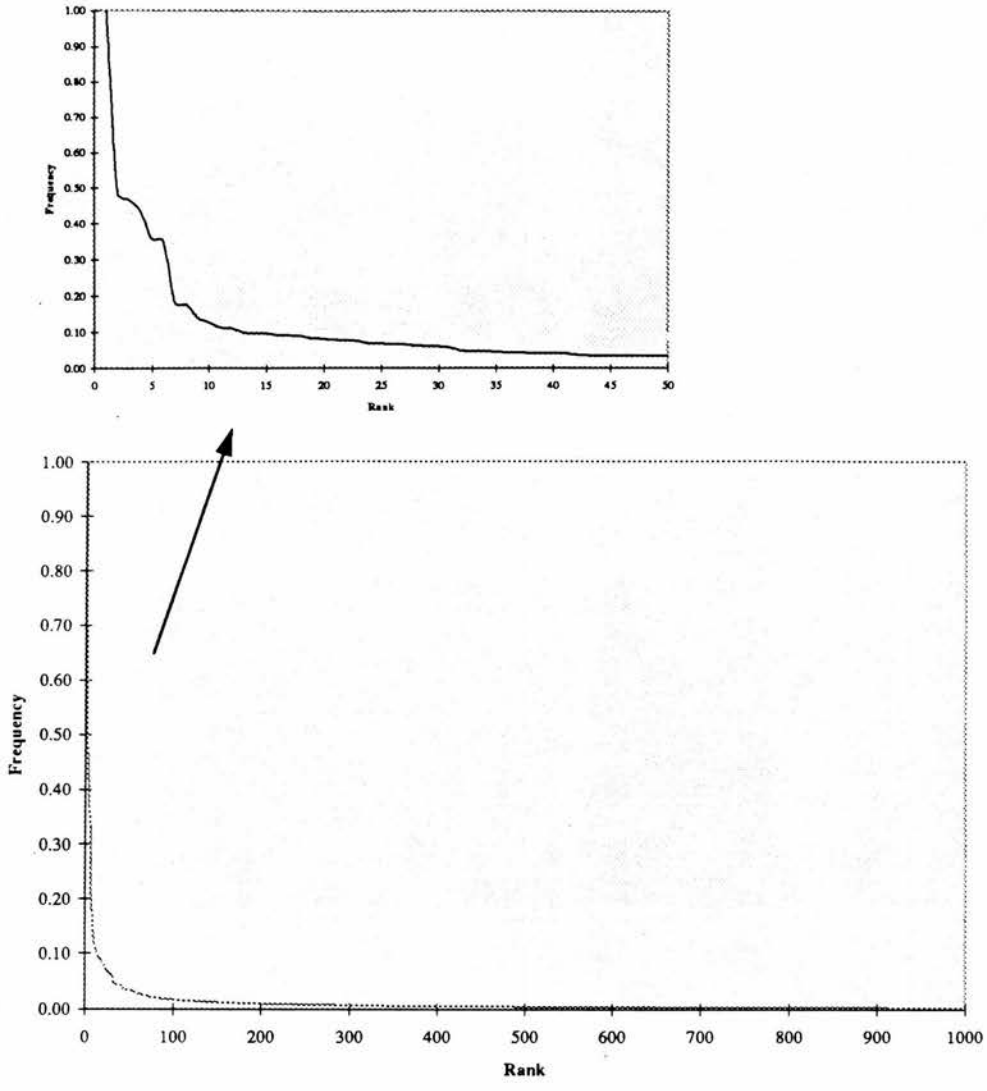
"empirically, these shortened vectors provide similar results to the full-length vectors, while being much easier to work with (p661)".

Their approach would only be appropriate if the size of the corpus used was accordingly large.

Lund et al. do not give the exact frequencies of the words they considered, but we can attempt to estimate these in approximate fashion from the frequencies of words used in this thesis. In the work to be described in Chapter 4, frequency information was calculated for the most frequent 1,000 words occurring in a 10,000,000 word sample of the Wall Street Journal. The result is shown (with an inset for clarity, showing data for the most frequent 50 words only) in figure 3.2.

[4] Lund (*personal communication*) provided the author with a list of the 100 words whose column variance was largest in these sorts of analyses. An inspection of this list revealed that these were indeed high frequency items.

Figure 3.2: Empirical Relationship Between Word Frequency and Rank in the Wall Street Journal Corpus



In this corpus, the most frequently occurring word is 'the', comprising 5.76% of all the word tokens it contains. As can be seen from figure 3.2 above, the frequencies of the unique words in the corpus do indeed appear to follow Zipf's law as stated in equation 3.1 above. Thus, if the most frequent word comprises 5.76% of all the word tokens, the second most frequent would comprise approximately 2.78% of the corpus, the third most frequent 1.82% of the corpus, and so on. Roughly speaking, we would expect the word frequencies in Lund et al.'s corpus to follow this sort of pattern. On the assumption that their most frequent word also occupied 5.76% of their corpus, we

can predict the approximate frequencies for the less frequent words. The 1000th most frequent word, according to Zipf's law, would occupy 0.0041% of the words in the corpus (although in the Wall Street Journal used here, the figure was actually found to be approximately 0.01%), which for a 160,000,000 word corpus would correspond to 6,560 occurrences. Approximately speaking, the 10,000th most frequent word would occupy 0.00036% of the words (576 occurrences), the 50,000th most frequent word would occupy 0.000067% of the words (107.2 occurrences), and the 70,000th most frequent word would occupy 0.000047% of the words in the corpus (75.2 occurrences). Clearly, these are extremely small proportions of the corpus, and are also changing very little as we look at successively less frequent words. The 40,000 words between the 10,000th most frequent and the 50,000th most frequent words in the corpus only causes a difference in the percentage of the corpus occupied of 0.000293 percentage points; the slope of the curve here is so flat that for each extra word considered, the reduction in the percentage of the corpus occupied is on average only 0.000000007325 percentage points. Between the 50,000th most frequent word and the 70,000th most frequent word, the difference in the percentage of the corpus occupied is only reduced by 0.00002 percentage points; per extra word considered, the reduction in the percentage of the corpus occupied is only 0.000000001 percentage point. Given the extremely shallow gradients here, it is not surprising that the variance was found to be so low across columns of Lund et al.'s matrix which reflected the bigram frequencies of many of the less frequent words. Of course, these figures are all estimates based on a different corpus, but the pervasive nature of Zipf's law suggests that they will not be likely to differ greatly from the true situation.

It seems to be a reasonable conclusion, based on these estimates, that Lund et al. could have saved themselves a considerable amount of time and computational resources by reducing the number of words being considered from the outset.

Notwithstanding these criticisms, Lund et al. (1995) did eventually end up with a number of empirically derived vectors for words in their corpus. A reduction of the 200-dimensional vector space to 2 dimensions using a multidimensional scaling procedure revealed that words with similar meanings tended to be close together in

the 2 dimensional space. It was possible to differentiate three groups of words in this way, corresponding to geographic regions ('china', 'europe', 'russia', 'america', and so on), animals ('kitten', 'dog', 'puppy', 'mouse', and so on), and body parts ('head', 'face', 'ear', 'nose', and so on). They then sought correlations between the inter-vector distances in semantic space and the psychological measure of semantic priming.

Lund et al. found that related words (words located close together in the space) produced a significant priming effect, when compared with unrelated words. They found, furthermore, a priming effect both for semantically-related items, and for associatively-related items. However, following analysis of the relationships between the vectors of particular types of stimuli and the performance of human subjects, it was concluded that the vectors resulting from the distributional analysis were more semantic than associative in nature. This was taken to suggest that first-order associations amongst words, which are associations due to the temporal ordering of words in the language data, are a less important part of structural semantics than second-order associations, which reflect the patterns of intercorrelations amongst word *use* within the corpus. Lund et al. felt that the vectors obtained from the corpus reflected the latter type of association more than the former. Lund et al. also supported the idea that second-order associations were an important aspect of the vectors being used by appealing to the sort of replacement criterion also referred to by Finch and Chater (1992a, 1992b). They made the assertion here that words which are semantically similar can be interchanged within a sentence, whilst words which are purely associatively related produce awkward sentences when interchanged in this way. For example, in the sentences 'the child slept on the bed', and 'the child slept on the table', 'bed' and 'table' are semantically related items which can be comfortably interchanged. However, in the sentences 'the child slept in the cradle' and 'the child slept in the baby', 'cradle' and 'baby' are associatively related items which clearly cannot be so legitimately interchanged. Thus, Lund et al. make the important claim that

"the semantic vectors take us beyond simple co-occurrence in that they are really measures of context (p664)".

Bullinaria and Huckle (1996)[5] have also investigated the relationship between priming effects and the similarity of high-dimensional vector representations for words. A cascaded feed-forward neural network, trained by a form of the back-propagation algorithm (Rumelhart, Hinton, and Williams (1986)) was used. The task of the network was to map 270 monosyllabic words from orthography to semantics. Rather than using a binary representation for the semantics, these were provided in the novel form of statistical vectors derived from the Wall Street Journal. The vectors were initially of 400 dimensions, but were reduced to 30 dimensions by projecting them onto the 30 dimensional sub-space containing the maximum variance.

Semantic priming was modelled in two ways. Firstly, the time was recorded for the network to settle into a stable output semantic state once it had been provided with the orthography of a word at the input units. This was termed 'settling time'. Secondly, the time was recorded for consistency checking; that is, the time taken for activation to flow from the input layer to semantics and back to phonology again. This was termed 'consistency time'. The performance of the network was assessed by comparing the 'reaction time' for each of the words considered when primed by the three closest words in semantic space with that when they were primed by the three furthest words. Both forms of simulated reaction time were found to show significantly faster times for the near sets of words than for the remote ones, confirming that a form of semantic priming was indeed occurring. However, these results were noisy and not in the predicted direction in every case. Noise is, nonetheless, a characteristic of human priming data. It was concluded that further simulations with different words and networks with different starting weights would be appropriate in further exploration of the phenomena observed in the neural network used here.

[5]Bullinaria is at the Centre for Speech and Language, Department of Psychology, Birkbeck College, London. In the work described in Bullinaria and Huckle (1996), Bullinaria wrote the neural network simulation software and carried out the simulations. Huckle constructed the semantic vectors through a statistical analysis of the Wall Street Journal and wrote the software to carry out this part of the work. The manuscript is included in Appendix D.

3.3 Conclusions

We have now seen that a considerable body of statistical language work has been built up by researchers in Computational Linguistics, in addition to a much more limited body of work carried out by psychologists and Cognitive Scientists. This work has often been concerned with the problem of categorizing words into coherent syntactic groups, using statistical information. Various different methods have been employed, and we have seen that the results obtained are sufficiently impressive to suggest that the statistical approach is far from irrelevant. Nonetheless, we have examined a recent approach to the problem of semantic categorization in depth and have noted that caution needs to be exercised when conducting large-scale statistical analyses of this type.

Having observed that there are useful statistical techniques which might be used for furthering our aim of investigating the extent to which statistical information can be useful in categorizing word meanings, we can now progress to the stage of designing and carrying out relevant analyses.

4. THE 'STANDARD' STATISTICAL ANALYSES

4.1 Using Vector Representations for Words

In accordance with the aims discussed in Chapter 2 of this thesis, we would like to be able to find a way to represent words which does not make any *a priori* assumptions about them. Since we are intending to investigate the performance of a system which learns in an unsupervised fashion on the basis of the statistical structure within the language data, it would be undesirable, for example, to have to specify in advance whether a particular word is a noun or a verb, or is abstract or concrete. Instead we desire a method for representing words which is uniform for all words.

Such a method is provided by the use of statistical context vectors to represent the words being encountered, as we have seen in Chapter 3. Grefenstette (1993) refers to methods of this type as "knowledge-poor" methods, since they do not require any prior semantic information and depend on the frequency of co-occurrence of words to determine similarity. As to the distinction between supervised and unsupervised learning, we may regard such methods as satisfying our requirement of being unsupervised since we supply no information about 'desired' or 'correct' responses at any point[6]. In informal terms, this sort of approach represents each word w_i by a vector, each of whose components says something about the statistical behaviour of a second word w_j in relation to the word being represented. As we noted in Chapter 1, for ease of reference we shall refer to the word w_i being represented as the 'target word', and the words w_j contained within the components of the vector as the 'context words'.

With this type of approach we are, of course, restricting our attention to the use of intralinguistic information in developing a semantic categorization of language. In the case of human acquisition of such a categorization, extralinguistic information presumably plays an important role. By focussing on information which is internal to

[6] Pereira, Tishby, and Lee (1993) argue that the collection of bigram statistics is a type of learning which falls between the supervised and unsupervised types, since the words whose bigram statistics are being learned do not themselves have any internal structure.

the language, however, we can attempt to satisfy our aim of investigating the extent to which the use of such information is likely to be important. We need not, therefore, make assumptions about the way in which extralinguistic information is exploited. As Redington, Chater and Finch (1993) point out, representing such information would in any case be highly problematic:

"Whilst language external factors are presumably of considerable importance, they are very difficult to model computationally, given our almost complete lack of knowledge as to how they can be appropriately represented. Empirical data concerning the child's representation of the world remains both anecdotal in nature, and difficult to interpret (p849)"

For each target word, then, we represent in the form of a vector the statistical behaviour of a set of context words in relation to it. When we consider more than a small number of context words, we assume that the vectors used to represent any two target words will not be identical. We assume further, and importantly, that target words which are intuitively similar in meaning will have context vectors which somehow reflect this similarity.

Once each word has been represented by a statistical vector, we need to put into practice our assumption that the similarity of the vectors will indeed reflect the similarity in meaning between the words being considered. The approach taken here towards achieving this, as with several of the pieces of research discussed in Chapter 3, is to firstly calculate some metric of distance between all pairs of vectors being considered. Once this step has been completed, we then proceed to pass the resulting matrix of distances to a hierarchical cluster analysis procedure which can provide a visual representation of the similarities and differences between the vectors.

As Charniak (1993) has pointed out, it is not yet clear from distributional language research which distance metric is the most appropriate one to use in analyses of this kind. Different metrics are consequently encountered in different pieces of work; Finch and Chater (1992a, 1992b), for example, have favoured the use of the Spearman Rank Correlation coefficient (because it appeared to give the best results), whilst Brown, Della Pietra, deSouza, Lai, and Mercer (1992) used the information theoretic measure of mutual information. In the analyses to be described in this chapter, two metrics were used. Firstly, a simple Euclidean distance measure was used, and secondly the Spearman rank correlation coefficient.

4.2 Methodology of the Analyses

4.2.1 The Moving Window

Statistical information about the language being considered is gathered by the system by 'reading' a text from a large corpus. In common with other similar approaches, a 'moving window' is used during this process. When each target word is encountered, the window pauses with the target word at its centre and those neighbouring words which are contained within the window are examined. The frequency count for each of the context words which appears within the window is incremented and stored. The window then moves to the right until it encounters the next target word, at which point the window pauses again and the procedure is repeated.

In the analyses to be described various windows of various lengths were compared since there was no firm basis on which to prefer a particular window length. In each case, the window extended both sides of the target word. The psychological plausibility of using right-sided context is debatable, but since useful statistical information is likely to be obtained through its use, and because it was not initially clear how much information the system would require in order to work at all, the decision was made to include it. This issue is discussed further in Chapter 9.

The analyses described here, which we shall refer to as the 'standard' analyses, are similar to those reported by Finch and Chater (1992a, 1992b), but differs in that the aim here is to explore semantic, rather than syntactic categorization, and in that we do not record information about the ordering of the context words. This latter difference was justified on the basis that, whilst it may well be reasonable to assume that human beings have knowledge of the order in which context items occur, it would be preferable to incorporate as few *a priori* assumptions into the system as possible.

4.2.2 Vector Components

In the analyses carried out, each component j of the context vector represents the probability $p(w_j|w_i)$ that any one word position in a 'context window' will be occupied by a context word w_j, given that this window is centred on word w_i.

The reason for this approach is that initially we would like to do nothing more complicated than to record, for each target word w_i, the frequency with which each context word w_j occurs within a certain distance of it. The distance is of course determined by the moving window centered on w_i. However, we also need to take into account the frequency of the target word itself because this will inevitably affect the frequency counts for the context words and consequently may distort the results. The frequency counts for context words in a vector representing an infrequent target word will inevitably be lower than frequency counts for the same context words in a vector representing a frequent target word. Depending on the distance metric used, this may obscure any fundamental similarity between the vectors. For instance, it might be that the target words 'camel' and 'dromedary' have very similar statistical contexts in the English language, but a straightforward frequency count for the context words near them may fail to capture this because one of these words occurs much more frequently than the other.

To deal with this difficulty, we simply normalize the frequency count for each context word in a vector by dividing by the frequency of the target word for that vector. This enables us to express the frequency counts for the context words relative to a single occurrence of each target word. In other words, the frequency counts become the *probability* that, given that we are looking at a particular target word, a particular context word will occur with a certain distance of it.

4.2.3 Distance Metrics

The distance metrics used in the analyses carried out here were Euclidean distance and the Spearman Rank Correlation coefficient.

The Euclidean distance $d(\mathbf{x},\mathbf{y})$ between two vectors $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ is given by:

$$d(\mathbf{x},\mathbf{y}) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} \qquad (4.1)$$

The Spearman Rank Correlation coefficient r_s between two vectors $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ is given by:

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)}, \qquad (4.2)$$

where d is the difference between the ranks assigned to the members of each pair of corresponding dimensions in the two vectors.

The Spearman Rank Correlation coefficient is a fairly crude measure of similarity in that it pays attention only to the ordering of the components in each vector; it is only necessary for the components in each vector to increase or decrease monotonically for there to be a correlation of 1 or -1. The Euclidean distance metric, on the other hand, is affected also by the shape of the distributions formed by the sets of components being compared.

4.2.4 Hierarchical Cluster Analysis

For each target word being considered, we create a vector of conditional probabilities as described above. Following this, the distance between each pair of vectors is calculated using one of the two distance metrics. This results in a matrix of inter-word distances, in which the distance between all pairs of vectors is contained. This in turn provides the input to a hierarchical cluster analysis procedure.

Cluster analysis is a descriptive statistical technique which can be used to provide a visual representation of the inter-word distances. It begins by discovering the shortest distance between any two words and by plotting them together as two neighbouring leaves on a tree-like diagram referred to as a dendrogram. The two words are then combined. The shortest distance between any two entities is then recalculated (this could be the distance between two words or that between a word and the new combined entity). This procedure repeats until all words have been considered and a complete dendrogram has been created.

There are many different ways in which the procedure can be carried out (see Lorr (1983) and Murtagh (1993) for reviews of these) and it has yet to be established which of these is the most appropriate in analyses of the present kind[7]. Indeed, it is frequently the case in research reports that reference is not made to the particular type of cluster analysis being carried out.

In the analyses to be presented in this chapter, a centroid method is used in the cluster analysis procedure. This particular method was chosen for convenience because pre-existing UNIX software which uses it could readily be adapted by the author to analyse the output of the programs written to read the text. The 'clusters' program which was adapted for use here was originally written and made generally available by Andreas Stolcke[8], and was obtained electronically. The centroid method starts, as we described above, by assuming that each of the target words belongs to its own cluster, and proceeds by iteratively merging these clusters into larger ones. At each time step, the algorithm looks for the smallest distance (either Euclidean distance or the Spearman Rank Correlation coefficient) between any two clusters and then merges these clusters together. When calculating the distance between one cluster and another, each cluster is represented by its centroid, which is the average of all the data points it includes. The process repeats until only one cluster remains.

---

[7] Future research in this area would benefit from a systematic study of this issue. As Charniak (1993) has pointed out, "... there is no doubt a lot of useful work to be done on choosing the proper metric and algorithm for clustering (p136)".

[8] Andreas Stolcke is at the Speech Technology and Research Laboratory, Menlo Park, California.

4.3 Results of the Analyses

Having described the details of the method used to obtain the statistical vectors, we shall now present a listing of the results obtained by analyses carried out using different combinations of settings for the window length and distance metric parameters. In the absence of firm reasons for particular choices in setting these parameters, their merits are explored empirically. In Chapter 5, an objective method for comparing the results obtained is presented.

4.3.1 Analysis 1

In this statistical analysis, the relevant parameter settings are given in table 4.1 below.

Table 4.1: Parameters Used in Analysis 1

Corpus	Wall Street Journal (1988/1989)
Number of Words in Corpus	9999402
Window Length[9]	1
Number of Target Words Considered	1000
Number of Context Words Used	1000
Distance Metric	Spearman Rank Correlation Coefficient

Here, each of the 1000 target words is represented by a context vector, as described above, which contains statistical information about the behaviour of the same 1000 words relative to the target words. The use of as many context words as this could be challenged on the basis that, in accordance with Zipf's law, some of these words will occur only infrequently in the text and may consequently introduce some unreliable statistics into the vectors. This issue was discussed at length in Chapter 3. When the programs to carry out these analyses were written, however, it was not clear how many context words *would* be appropriate for inclusion in the vectors. In the absence of any firm guidance here, it was decided to represent the target words 'in terms of themselves' in all the analyses to be carried out. That is, the same set of words would be used as target words and context words. A clear disadvantage of this approach is that as the set of target words increases in size, so will the number of unreliable context words being used. After the software had been written and the present analyses had been carried out, the work of Hughes (1994) became available. This

[9] Here, and elsewhere, the window length refers to the number of words enclosed within the moving window *on each side* of the target word.

work, which is discussed further in Chapter 5, reports an empirical evaluation of various clustering techniques suggesting that a context of only 100 lexical items can yield very good results, and that there is little to be gained from using more context words than this.

Despite the possibly non-optimal approach here, the results obtained showed clear evidence of a large number of groups of words which are intuitively related in meaning. These results will firstly be presented by giving the nearest neighbours for each of a number of target words. The nearest neighbours are established in this particular case using the Spearman Rank Correlation coefficient as the distance metric. Rather than giving nearest neighbours for all the 1000 target words analysed, attention is restricted to 50 target words chosen *at random* from the full set of target words. This method of presenting results was used by Schütze (1993a). The 10 nearest neighbours for each of these 50 target words, along with the relevant correlation coefficients, are given in table 4.2.

Table 4.2: Nearest Neighbours for the Target Words Considered in Analysis 1 (Spearman Distance Metric, Window Length=1)

Target Word	10 Nearest Neighbours (Spearman Correlation Coefficient)
able	kind (0.801) composite (0.797) lot (0.773) reporter (0.759) trying (0.753) done (0.749) looking (0.749) try (0.748) thought (0.743) bon (0.742)
above	below (0.721) composite (0.643) lot (0.631) unchanged (0.621) tons (0.620) kind (0.619) percentage (0.610) counter (0.607) able (0.604) preferred (0.594)
analyst	composite (0.697) reporter (0.684) minister (0.652) lot (0.652) kind (0.644) spokesman (0.644) able (0.642) smith (0.641) attorney (0.640) counter (0.640)
base	lot (0.660) composite (0.651) kind (0.643) reporter (0.638) reason (0.632) role (0.630) mergers (0.629) tender (0.619) bankruptcy (0.619) percentage (0.616)
close	kind (0.630) begin (0.622) done (0.614) reason (0.610) try (0.608) effect (0.606) fact (0.605) closely (0.600) able (0.599) lot (0.598)
concern	company (0.713) maker (0.641) subsidiary (0.641) division (0.619) concerns (0.617) firm (0.617) unit (0.602) group (0.592) businesses (0.587) units (0.586)
deal	kind (0.694) lot (0.685) fight (0.685) thing (0.683) reason (0.682) question (0.669) effort (0.658) know (0.651) thought (0.648) number (0.647)
despite	while (0.569) during (0.558) after (0.554) in (0.545) following (0.539) because (0.537) as (0.534) compared (0.532) although (0.529) but (0.520)

Table 4.2 (contd.)

even	it (0.589) you (0.572) but (0.569) so (0.561) that (0.556) only (0.552) still (0.550) there (0.547) just (0.537) lot (0.533)
expects	expect (0.628) owns (0.612) posted (0.605) wants (0.605) composite (0.599) lot (0.597) reduce (0.590) bnk (0.588) fin (0.578) gained (0.578)
family	kind (0.636) children (0.626) reporter (0.624) lot (0.617) men (0.606) fight (0.595) closely (0.592) thing (0.591) fact (0.589) women (0.586)
gained	posted (0.707) dropped (0.688) closed (0.673) rose (0.670) fell (0.665) lost (0.663) composite (0.663) declined (0.650) acquired (0.644) owns (0.642)
general	composite (0.543) secretary (0.513) jones (0.513) st (0.512) jr (0.509) dow (0.508) de (0.504) los (0.500) san (0.499) kind (0.497)
george	composite (0.671) paul (0.670) reporter (0.654) jr (0.648) kind (0.647) lot (0.646) secretary (0.644) able (0.642) smith (0.636) justice (0.629)
germany	lot (0.703) reporter (0.694) europe (0.687) kind (0.686) tons (0.684) plc (0.684) thing (0.678) fight (0.677) minister (0.674) counter (0.673)
hard	kind (0.703) difficult (0.693) done (0.686) lot (0.686) able (0.670) bad (0.662) clear (0.659) thing (0.659) taken (0.658) doing (0.657)
included	composite (0.611) able (0.591) unchanged (0.582) completed (0.582) scheduled (0.581) percentage (0.579) kind (0.578) lot (0.575) reporter (0.570) tons (0.570)
independent	composite (0.630) publishing (0.621) reporter (0.620) closely (0.620) mergers (0.617) lot (0.612) kind (0.612) bankruptcy (0.610) able (0.602) role (0.594)
index	composite (0.640) tons (0.604) futures (0.585) kind (0.584) october (0.582) lot (0.580) sharply (0.578) yen (0.577) able (0.564) september (0.564)
it's	that's (0.644) he's (0.615) they're (0.610) i'm (0.593) we're (0.577) is (0.576) there's (0.562) so (0.535) there (0.529) he (0.516)
labor	bankruptcy (0.584) fin (0.580) bnk (0.577) mon (0.574) bon (0.572) composite (0.572) monetary (0.570) hong (0.567) lot (0.562) los (0.558)
making	trying (0.579) doing (0.575) using (0.569) getting (0.564) taken (0.561) looking (0.560) kind (0.560) taking (0.559) able (0.558) going (0.553)
men	women (0.726) children (0.721) thing (0.699) kind (0.697) lot (0.697) reporter (0.679) people (0.674) fact (0.671) me (0.670) leaders (0.668)
night	reporter (0.712) kind (0.712) thing (0.710) lot (0.705) question (0.697) done (0.696) reason (0.694) fight (0.684) fact (0.684) effect (0.681)
nuclear	composite (0.699) lot (0.664) hong (0.661) monetary (0.655) kind (0.651) de (0.646) able (0.646) bankruptcy (0.645) mergers (0.643) tons (0.639)
old	jr (0.565) reporter (0.555) lot (0.549) kind (0.545) attorney (0.520) thing (0.518) composite (0.509) head (0.509) young (0.509) justice (0.507)
operating	fourth (0.570) composite (0.565) percentage (0.563) net (0.560) fiscal (0.560) publishing (0.560) able (0.554) lot (0.553) posted (0.551) mergers (0.541)
paid	bought (0.637) done (0.635) thought (0.629) offered (0.626) taken (0.625) lost (0.625) seen (0.618) able (0.617) decided (0.612) kind (0.612)
partners	kind (0.641) directors (0.638) reporter (0.632) mergers (0.631) acquisitions (0.630) jr (0.627) plants (0.620) minister (0.617) lot (0.617) leaders (0.616)
percentage	composite (0.778) tons (0.744) lot (0.730) able (0.723) kind (0.719) monetary (0.702) hong (0.702) reporter (0.700) counter (0.695) los (0.694)
political	economic (0.598) democratic (0.586) composite (0.576) legal (0.574) anti (0.572) monetary (0.559) role (0.556) kind (0.556) bankruptcy (0.547) difficult (0.540)
preferred	composite (0.733) lot (0.711) kind (0.701) tender (0.694) reporter (0.688) mergers (0.677) bankruptcy (0.677) tons (0.674) able (0.674) series (0.665)
product	kind (0.608) store (0.607) products (0.604) reason (0.603) thing (0.595) software (0.593) number (0.590) form (0.585) effort (0.584) lot (0.581)
products	equipment (0.645) systems (0.625) parts (0.614) goods (0.611) services (0.605) product (0.604) computers (0.602) programs (0.596) stores (0.592) lines (0.590)
same	composite (0.561) single (0.545) counter (0.539) bon (0.528) difficult (0.526) kind (0.523) lot (0.522) fin (0.519) different (0.518) percentage (0.518)
should	will (0.778) would (0.776) could (0.759) can (0.751) must (0.723) might (0.698) won't (0.686) may (0.660) didn't (0.659) can't (0.628)
take	try (0.664) find (0.660) want (0.649) get (0.646) hold (0.638) taken (0.637) know (0.636) give (0.636) need (0.634) look (0.631)
they're	he's (0.671) i'm (0.635) we're (0.614) lot (0.613) it's (0.610) able (0.597) kind (0.595) that's (0.585) try (0.584) composite (0.580)
times	tons (0.643) kind (0.638) lot (0.638) weeks (0.633) done (0.631) fact (0.622) reason (0.612) thing (0.611) able (0.609) closely (0.607)

Table 4.2 (contd.)

transaction	amount (0.659) kind (0.657) acquisition (0.651) settlement (0.650) lot (0.648) reason (0.648) tender (0.647) deal (0.643) filing (0.643) proposal (0.639)
transportation	composite (0.665) reporter (0.632) de (0.628) publishing (0.625) lot (0.624) san (0.623) fin (0.617) able (0.615) bon (0.615) los (0.612)
use	believe (0.634) know (0.628) try (0.624) find (0.622) kind (0.617) receive (0.613) role (0.611) lot (0.610) reason (0.609) fight (0.600)
using	trying (0.622) doing (0.607) able (0.601) composite (0.595) kind (0.593) looking (0.585) lot (0.579) getting (0.578) making (0.569) seeking (0.563)
wall	hong (0.522) composite (0.515) san (0.509) last (0.502) merrill (0.492) compared (0.492) holds (0.483) percentage (0.479) los (0.478) bankruptcy (0.477)
wednesday	monday (0.701) tuesday (0.692) friday (0.687) composite (0.648) lot (0.625) sept (0.623) unchanged (0.618) tons (0.618) august (0.602) september (0.602)
week	month (0.670) weeks (0.660) year (0.634) months (0.619) night (0.616) ago (0.597) days (0.587) example (0.584) comment (0.582) december (0.577)
weren't	aren't (0.702) were (0.618) are (0.615) able (0.594) composite (0.591) counter (0.582) continue (0.582) try (0.575) lot (0.573) kind (0.566)
will	would (0.883) could (0.846) can (0.783) should (0.778) won't (0.754) must (0.740) might (0.726) may (0.724) didn't (0.700) wouldn't (0.656)
william	john (0.640) george (0.628) composite (0.622) richard (0.621) robert (0.619) james (0.618) david (0.614) michael (0.610) paul (0.607) de (0.602)
workers	employees (0.641) men (0.624) children (0.615) jobs (0.609) leaders (0.605) women (0.603) plants (0.596) customers (0.596) members (0.593) kind (0.584)

The results of analysing the data using cluster analysis are also presented. The full dendrogram is reproduced in appendix A, figure A.1, with the 50 random words indicated in capital letters. The groups of words present in the dendrogram differ from those which result from the straightforward nearest neighbours analysis above. As we noted earlier, this is because in cluster analysis, words are in general added to an existing cluster of words on the basis of their distance from an *average* of the words in that cluster.

4.3.2 Analysis 2

In this statistical analysis, the relevant parameters are given in table 4.3 below.

Table 4.3: Parameters Used in Analysis 2

Corpus	Wall Street Journal (1988/1989)
Number of Words in Corpus	9999402
Window Length	1
Number of Target Words Considered	1000
Number of Context Words Used	1000
Distance Metric	Euclidean Distance

The table containing the list of the 50 randomly chosen target words and their 10 nearest neighbours for this analysis is given in Appendix A, table A.2. The relevant dendrogram is shown in Appendix A, figure A.2

4.3.3 Analysis 3

In this statistical analysis, the relevant parameters are given in table 4.4 below.

Table 4.4: Parameters Used in Analysis 3

Corpus	Wall Street Journal (1988/1989)
Number of Words in Corpus	9999402
Window Length	2
Number of Target Words Considered	1000
Number of Context Words Used	1000
Distance Metric	Spearman Rank Correlation Coefficient

The table containing the list of the 50 randomly chosen target words and their 10 nearest neighbours for this analysis is given in Appendix A, table A.3. The relevant dendrogram is shown in Appendix A, figure A.3.

4.3.4 Analysis 4

In this statistical analysis, the relevant parameters are given in table 4.5 below.

Table 4.5: Parameters Used in Analysis 4

Corpus	Wall Street Journal (1988/1989)
Number of Words in Corpus	9999402
Window Length	2
Number of Target Words Considered	1000
Number of Context Words Used	1000
Distance Metric	Euclidean Distance

The table containing the list of the 50 randomly chosen target words and their 10 nearest neighbours for this analysis is given in Appendix A, table A.4. The relevant dendrogram is shown in Appendix A, figure A.4.

4.3.5 Analysis 5

In this statistical analysis, the relevant parameters are given in table 4.6 below.

Table 4.6: Parameters Used in Analysis 5

Corpus	Wall Street Journal (1988/1989)
Number of Words in Corpus	9999402
Window Length	5
Number of Target Words Considered	1000
Number of Context Words Used	1000
Distance Metric	Spearman Rank Correlation Coefficient

The table containing the list of the 50 randomly chosen target words and their 10 nearest neighbours for this analysis is given in Appendix A, table A.5. The relevant dendrogram is shown in Appendix A, figure A.5.

4.3.6 Analysis 6

In this statistical analysis, the relevant parameters are given in table 4.7 below.

Table 4.7: Parameters Used in Analysis 6

Corpus	Wall Street Journal (1988/1989)
Number of Words in Corpus	9999402
Window Length	5
Number of Target Words Considered	1000
Number of Context Words Used	1000
Distance Metric	Euclidean Distance

The table containing the list of the 50 randomly chosen target words and their 10 nearest neighbours for this analysis is given in Appendix A, table A.6. The relevant dendrogram is shown in Appendix A, figure A.6.

4.3.7 Analysis 7

In this statistical analysis, the relevant parameters are given in table 4.8 below.

Table 4.8: Parameters Used in Analysis 7

Corpus	Wall Street Journal (1988/1989)
Number of Words in Corpus	9999402
Window Length	10
Number of Target Words Considered	1000
Number of Context Words Used	1000
Distance Metric	Spearman Rank Correlation Coefficient

The table containing the list of the 50 randomly chosen target words and their 10 nearest neighbours for this analysis is given in Appendix A, table A.8. The relevant dendrogram is shown in Appendix A, figure A.8.

4.3.8 Analysis 8

In this statistical analysis, the relevant parameters are given in table 4.9 below.

Table 4.9: Parameters Used in Analysis 8

Corpus	Wall Street Journal (1988/1989)
Number of Words in Corpus	9999402
Window Length	10
Number of Target Words Considered	1000
Number of Context Words Used	1000
Distance Metric	Euclidean Distance

The table containing the list of the 50 randomly chosen target words and their 10 nearest neighbours for this analysis is given in Appendix A, table A.8. The relevant dendrogram is shown in Appendix A, figure A.8.

4.3.9 Analysis 9

In this statistical analysis, the relevant parameters are given in table 4.10 below.

Table 4.10: Parameters Used in Analysis 9

Corpus	Wall Street Journal (1988/1989)
Number of Words in Corpus	9999402
Window Length	25
Number of Target Words Considered	1000
Number of Context Words Used	1000
Distance Metric	Spearman Rank Correlation Coefficient

The table containing the list of the 50 randomly chosen target words and their 10 nearest neighbours for this analysis is given in Appendix A, table A.9. The relevant dendrogram is shown in Appendix A, figure A.9.

4.3.10 Analysis 10

In this statistical analysis, the relevant parameters are given in table 4.11 below.

Table 4.11: Parameters Used in Analysis 10

Corpus	Wall Street Journal (1988/1989)
Number of Words in Corpus	9999402
Window Length	25
Number of Target Words Considered	1000
Number of Context Words Used	1000
Distance Metric	Euclidean Distance

The table containing the list of the 50 randomly chosen target words and their 10 nearest neighbours for this analysis is given in Appendix A, table A.10. The relevant dendrogram is shown in Appendix A, figure A.10.

4.3.11 Analysis 11

In this statistical analysis, the relevant parameters are given in table 4.12 below.

Table 4.12: Parameters Used in Analysis 11

Corpus	Wall Street Journal (1988/1989)
Number of Words in Corpus	9999402
Window Length	100
Number of Target Words Considered	1000
Number of Context Words Used	1000
Distance Metric	Spearman Correlation Coefficient

The table containing the list of the 50 randomly chosen target words and their 10 nearest neighbours for this analysis is given in Appendix A, table A.11. The relevant dendrogram is shown in Appendix A, figure A.11.

4.3.12 Analysis 12

In this statistical analysis, the relevant parameters are given in table 4.13 below.

Table 4.13: Parameters Used in Analysis 12

Corpus	Wall Street Journal (1988/1989)
Number of Words in Corpus	9999402
Window Length	100
Number of Target Words Considered	1000
Number of Context Words Used	1000
Distance Metric	Euclidean Distance

The table containing the list of the 50 randomly chosen target words and their 10 nearest neighbours for this analysis is given in Appendix A, table A.12. The relevant dendrogram is shown in Appendix A, figure A.12.

4.4 Discussion

The dendrograms and tables of nearest neighbours corresponding to the analyses carried out reveal that the results obtained are not random groupings of words, but are often intuitively familiar ones which do indeed share similarity in meaning. Whilst noise is also evident amidst these groupings, the structures which can be identified confirm that the use of statistical information alone has been a useful source of information regarding the similarities and differences between the words in the corpus.

The results which have been presented here are of a descriptive nature, and, as we shall discuss further in Chapter 5, it is not desirable to go to great lengths in analysing them without a more objective means of assessment. However, it is of interest to examine some of the main features of the results obtained.

Even when using the very shortest window length of 1 word either side of the target word, interesting groupings of words are present. Many of these, such as the collection of modal verbs and prepositions, are of a more syntactic than semantic nature. This, of course, is not surprising, since, as we have seen, analyses of a similar type have been applied to the problem of syntactic categorization with some success by other researchers. It is interesting to note, nonetheless, that inflected forms of a particular word are also sometimes grouped together, as in the case of the words 'include', 'included', 'including', and 'includes' (figure A.1). This is intuitively appealing, given the deeper semantic similarity between these words. Plural and singular nouns are, similarly, often grouped together.

Table 4.2 exhibits an idiosyncracy which appears not be present in the other tables of nearest neighbours; there are a small number of neighbours which are listed amongst the 10 nearest neighbours for *several* target words. Most notable amongst these neighbours is the word 'composite', with the word 'reporter' showing a similar

tendency. Since this phenomenon is not apparent in the other analyses, and since the analyses were run under identical conditions apart from the window length and distance metric parameter settings, it appears to be a genuine effect which arises from the use of a window length of 1 with the Spearman distance metric (it does not occur when using a window length of 1 with the Euclidean distance metric). Under these conditions, it would appear that the statistical vector for some words (such as 'composite') can come to be positioned in the high dimensional space such that they are in close proximity to several other words, when distance is measured using the Spearman coefficient.

More purely semantic groupings are also evident amongst the results obtained, with numerous examples of similar commodities, cities, nationalities, famous people, and so on, being grouped together. Many of these, since they are words which would not normally be expected to occur close together in the text, are likely to derive from similarity of a semantic, rather than an associative, nature. This sort of phenomenon provides particularly strong empirical support for the assumption that words that are similar with respect to their statistical context may also be similar with respect to meaning; in other words, the obtained results are not solely reliant on associative relationships between words. To take an example from figure A.3, one would not expect the words 'television' and 'tv' to occur within a small number of words of each other on a consistent basis; rather, one would expect them to be used more independently than this, but, of course, often occurring in very similar contexts - since the two words are very close synonyms. The fact that the two words are nonetheless grouped together suggests that this similarity in context has been detected and is sufficient to determine that the two words are closer to each other than to any other words in the corpus.

It is clear from the results obtained that antonyms are very often grouped together. In figure A.5, for example, 'up' and 'down' are placed closely together, as are 'loss' and 'gain', 'small' and 'huge', and 'buy' and 'sell'. Again, these are pairs of words which have, statistically speaking, very similar contexts. However, whilst we recognize the close relationship between the meanings of these words, we are nonetheless aware

that there is, of course, an important difference between them. The simple approach taken in this chapter appears not to be able to make this distinction in many cases. Although it would be interesting for future approaches to investigate what information must be obtained in order for antonyms to be distinguished, the fact that they are grouped together here is understandable, and, indeed, has been noted as a potential characteristic of child language acquisition. Clark (1979), for example, notes that even a child who is at an advanced stage in acquiring word meanings will know of a number of word pairs which share a large number of semantic features, but are nonetheless antonyms. Clark proposes that it is only when the polarity of one of the features in the representation is set correctly that the child will interpret the two words correctly, and reports experimental evidence which shows that children do initially confuse the two members of such word pairs.

Associative relationships become more prevalent as the window length used in the analyses is increased. This is to be expected, since the similarity between the contexts of nearby items increases as the window length becomes larger. In the case of a window length of 100 words each side of the target word, two neighbouring target words will have context windows which overlap completely apart from one word at each end of the window and the positions of the target words themselves. Thus 94% of the context being recorded for the two words is the same, and, in the absence of information being recorded about the ordering of the context words, this will of course mean that the two target words will be very close to each other in the high-dimensional semantic space. Even pairs of target words which are not immediate neighbours, but are rather more distant from one another, can share a high proportion of their contexts in this way. For these reasons, groupings such as 'wall', 'street', and 'journal' (figure A.10), 'real' and 'estate' (figure A.11), and 'chief' and 'executive' (figure A.12), occur with the longer window lengths. With shorter window lengths, of course, these sorts of effects occur less frequently; in the extreme case where a window length of 1 each side of the target word is used, neighbouring target words share no context at all.

4.5 Conclusions

In this chapter, 12 analyses using vector-based representations for target words have been conducted, exploring different settings for the distance metric and window length parameters. The results have been presented both in the form of tables of nearest neighbours and in the form of dendrograms produced by a hierarchical cluster analysis procedure. In each case, the results have provided evidence of rich structures, with the occurrence of both syntactic and semantic groupings of words.

Although these results are striking, they have been presented in a descriptive fashion only. It is desirable to achieve a more objective means of assessment, and this issue will be addressed in the next chapter.

5. *EVALUATING THE ANALYSES*

5.1 The Problem

In the preceding chapter, a number of distributional statistical analyses were presented. These explored the effects of altering certain parameters (distance metric and window length) while collecting statistics from a corpus. Following a stage in which cluster analysis was carried out, a dendrogram was produced for each analysis, providing a graphical depiction of the results of each analysis. For each analysis, a list of the ten nearest neighbours to a random sample of 50 target words was also presented.

Having conducted these analyses, we need to examine the results in order to try to provide some sort of answer to one of our original aims; that of ascertaining the extent to which simple statistical methods might provide a plausible means of developing a categorization for word meanings without external supervision. Ideally, it would be attractive to be able to arrive at some kind of objective measure for each analysis, which would indicate how well or how poorly that particular set of parameters had behaved in producing such a categorization. This would then allow us to perform comparisons between the analyses, giving some basis for preferring one particular set of parameters over another.

This, however, is not a straightforward matter. The dendrograms produced by cluster analysis are descriptive in nature and do not readily provide any objective score with which to evaluate the results obtained. It is possible to *compare* dendrograms with each other, in order to establish how similar or different they may be. Lapointe and Legendre (1995), for example, have described a non-parametric statistical test, known as the double permutation test, which attempts to do this. Such a test, however, says nothing about the appropriateness of the clustering in the dendrograms being compared. Where cluster analysis is used to conduct research in language classification, it has frequently been the case that dendrograms, or lists of clusters from dendrograms, are presented without an attempt at providing an objective measure of success. As Hughes (1994) notes:

"Immediately the question of deciding which of many clustering schema provide the best classification according to some measure is raised. This has largely been skipped over by researchers ... (p22)",

and Redington, Chater and Finch (1993) point out that:

"A quantitative measure of the 'goodness' of clustering/categorisation is obviously required. This would ideally allow comparison of results independently, to some extent, of sample size, and hopefully across languages (p853)".

Hughes goes on to describe the typical method used by researchers as the "*looks good to me*" approach:

"Evaluating a clustering is typically done by the programmer using a *looks good to me* approach. To an extent he/she can feel how good one clustering is over another because he/she has an intrinsic understanding of the processes that produced it. However, he/she also has a vested interest in making his/her program look good (Hughes 1994, p80)."

Similarly, Grefenstette (1993) states that:

"Evaluations of results produced ... are often ... limited to visual verification by a human subject or left to the human reader (p2)".

Finding an alternative to this intuitive method of evaluation, however, is not easy. First of all, some kind of benchmark is required. This benchmark is itself, however, unlikely to be based on anything other than intuition. We could, as Hughes also points out, ask an expert (such as a linguist) to evaluate a classification, but the evaluation will of course depend upon the particular assumptions and biases made by the expert.

Of course, the impossibility of avoiding the use of intuition in evaluating language classifications illustrates one of the main issues facing this thesis; as human beings, we are all able to develop and use a classification of word meanings, but we have little idea of how we achieved this and simply cannot provide rigourous definitions for the words that we use. Given this, the goal of trying to use a benchmark for evaluating classifications which is anything other than an intuitive one seems highly problematic.

By using the sort of benchmark provided by a single linguistic expert, we are assuming that the judgements of the expert will reflect the 'true' clustering that has been carried out by the human brain. As has already been noted, this runs the risk of becoming distorted by the particular characteristics or biases of the expert (though see Agarwal (1995) for an implementation of the approach).

There is also a practical problem facing evaluation by a linguistic expert; namely, the scale of the task required. In the analyses described in Chapter 4, 1000 words were presented in each case. Providing an evaluation for each of the analyses would be a formidable task, which would become impossible with larger numbers of analyses involving a larger number of words.

5.2 A Possible Compromise Solution

One possible way around the problem of finding a satisfactory benchmark with which to compare the results of different clustering methods, then, would be to use a benchmark which does not depend upon the decisions of a single expert, and to use one which could be automated to avoid the practical demands which would otherwise be placed on one individual.

Hughes (1994) has proposed one such solution for dealing with the need for a benchmark in *syntactic* clustering. In his scheme, the benchmark is provided by a version of the LOB corpus in which the words are tagged with a set of 23 syntactic tags. A 'benchmark clustering' was then obtained by clustering the words in the corpus on the basis of the similarity of the tags assigned to them. Thus two words which are always assigned the same tag in the LOB corpus would be clustered close together, while words would be clustered progressively further away from each other as the overlap between the types of tags assigned to them decreases. The benchmark clustering was found to produce a dendrogram in which the clusters could be identified as containing words of recognized syntactic classes, such as prepositions, nouns, articles, adverbs, and so on.

The evaluation could then be carried out by cutting the dendrogram resulting from some analysis into a number of sub-clusters. Each sub-cluster is labelled with one of the syntactic labels from the benchmark dendrogram. This is carried out on the basis of the label to which the majority of the words in the sub-cluster were assigned in the benchmark dendrogram. A score is then calculated by counting the number of words in the sub-cluster which are of the type indicated by its label.

One possible drawback to this scheme, as Hughes himself notes, is that the original tagging of the benchmark corpus will of course itself depend upon the assumptions of those who were responsible for doing the tagging. However, Hughes showed that it was of use in evaluating the appropriateness of a number of syntactic analyses. Using vectors of a similar type to those described in Chapter 4, Hughes found that the best performance was obtained when using a Manhattan distance metric and Ward's method for clustering.

Another approach using syntactic tags as a benchmark was adopted by Zavrel and Veenstra (1995). These authors used the parts of speech assigned to words in the annotated version of the Wall Street Journal as their benchmark. For each target word considered, the words clustered close to it were examined. The 'precision' of the neighbourhood of each target word was assessed by considering the proximity to the target word of neighbours assigned the same part of speech and the proximity of those assigned a different part of speech. Target words having high precision would have all neighbours with the same part of speech closer to them than neighbours with a different part of speech, whilst for lower precision words this would not be the case. Once the precision value had been calculated for all the target words, an objective measure had been obtained with which different analyses could be evaluated.

Redington, Chater, and Finch (1995) used a similar method in evaluating the success of syntactic analyses of text corpora. Using a corpus generated with a stochastic context-free grammar, they assessed the correspondence between the clusters in the dendrogram resulting from cluster analysis of the corpus and the original categories in the grammar. However, rather than reaching a score by counting up the number of matching items, as in the case of Hughes (1994), Redington et al. (1995) calculated the amount of mutual information between the categories in the grammar and those in the dendrogram, for each of the possible points at which the dendrogram might be cut. This procedure was also carried out on adult speech from the CHILDES corpus (MacWhinney and Snow, 1985), which is a real corpus containing transcribed speech from conversations between adults and children. The words in this corpus were given a canonical categorization based on the classifications assigned to them by the Collins Cobuild lexical database. It was shown for each of these corpora that the

dendrograms resulting from simple distributional analyses could be highly informative (relative to a chance permutation of the words) about the syntactic categories of the words involved.

Both of these approaches provide a benchmark which does not rely on the decisions of a single expert. In Hughes' case, the benchmark used was a tagged version of the LOB corpus, where the tagging was originally conducted by a team of individuals; in the case of Redington et al. (1995), the Collins Cobuild database provided the tags for the words considered. This database lists the frequency with which various syntactic tags were applied to the words in a large corpus of English. Both approaches are also capable of being carried out rapidly on a computer, and so the evaluation is perfectly practicable.

The work of Hughes (1994) and of Redington et al. (1995), however, is concerned chiefly with syntactic, rather than semantic, classification of natural language. For approaches of this kind to be useful in providing a reasonable benchmark for the sorts of analyses being addressed here, some kind of semantic benchmark is necessary. Since it is one of the assumptions of this thesis that no rigourous semantic definitions for most words can be provided, it follows that no benchmark corpus complete with 'semantic tags' is likely to be found. As was discussed earlier in the thesis, it seems more reasonable that words can only really be described in terms of other words which are more or less similar to them in meaning; this is essentially what a dictionary entry is. Indeed, Schütze and Pedersen (1993) have suggested that dictionary construction could be based upon vector based representations for the relationships between words.

Were it possible to carry out an exhaustive analysis on a corpus as large as all the language to which a typical human being has ever been exposed, we might be able to carry out a distributional analysis and produce a clustering which would capture the 'true' relationship between the meanings of the words we use. Quite apart from the obvious practical problems involved in using such a large amount of text, however, we also have the problem of having no idea *how* the large-scale distributional analysis

and clustering should be carried out in order to obtain the sort of conceptual structure used by human beings (always assuming, of course, that these procedures *can* potentially be used in some way to reach such a structure; this assumption may, of course, turn out to be far from true).

The best compromise in providing a semantic benchmark would, as before, seem to be that of finding a semantic classification of language which does not rely on the assumptions of a single individual, and which can be automated. But if no 'semantic tagging' is possible and if distributional methods cannot themselves be relied upon to produce a useful benchmark, then what might a reasonable solution be?

One readily available source of semantic classifications which presumably meets with widespread acceptance is a thesaurus. Roget's Thesaurus, in particular, has been used on occasion in statistical language research because it is available in electronic form.

A version of Roget's Thesaurus, which has been in use for many years, was used by Yarowsky (1992) as the basis for a system required to perform word-sense disambiguation. Word-sense disambiguation is a topic which will be discussed further in Chapter 6. It is, in brief, the task of deciding which of a number of possible senses of a word is appropriate, given the particular context in which a word appears.

In Roget's Thesaurus, each word falls into a number of categories. Yarowsky (1992) reasoned that since these categories tend to correspond to distinctions between the senses of each word, it would be useful to develop a system which could decide upon the most likely of these categories for a particular word when seen in context; such an approach might prove to be a useful means for selecting the appropriate sense of that word. This method was shown to be an effective one; Yarowsky applied the system to a 10,000,000 word corpus taken from Grolier's Encyclopedia and found that the system correctly disambiguated 92% of the occurrences of 12 polysemous words in this corpus. His approach is discussed in more detail in Chapter 6.

Brady (1993) has also used a version of Roget's Thesaurus within the context of representing concepts corresponding to word meanings. Brady sought to apply the notion of concept lattices to Roget's Thesaurus in order to improve the representation of knowledge within it. Concept lattices are a means of grouping objects and attributes into concepts, with the concepts organized in a lattice. In the resulting structure, some concepts are subconcepts of others.

Grefenstette (1993) used Roget's Thesaurus (and a version of Webster's dictionary) as a semantic benchmark (or 'gold standard') to evaluate the appropriateness of various methods of semantic classification, and his approach is thus of immediate interest for the present work. As Grefenstette notes,

"We would expect that any system claiming to extract semantics from text should find some of the relations contained in this resource (p2)"

Grefenstette sought to compare the success of a technique based on the use of a moving window, as used in the analyses presented in the preceding chapter, and a technique which used syntactic information to obtain semantic information about various words.

The syntactic method approached the problem of achieving a semantic categorization for words by extracting the syntactic context of each word in the corpus. The syntactic categories of the context words were defined using a computer based grammar, and the end result was that the context for each noun in the corpus reflected all the adjectives, nouns, and verbs which entered into syntactic relations with it.

The window based method, rather than using information about the syntactic context of the words being considered, took account only of the word tokens which occurred within 10 words either side of each target word and within the same sentence. Grefenstette's method restricted attention to context words which could be nouns, adjectives or verbs (according to a lexicon giving the possible parts of speech for each word). In its broad approach, then, this method is similar to that described in Chapter 4.

With each analysis, attention was restricted to the 2661 nouns which occurred 10 times or more in the corpus of 4,000,000 words taken from Grolier's Encyclopedia. In each case, the contexts for the nouns were compared for similarity using a weighted Jaccard measure, which gives a similarity measure between 0 and 1.

For each of the nouns considered, the noun calculated using the Jaccard measure as being most similar to it was recorded. Having found this 'nearest neighbour', Roget's Thesaurus was consulted to see whether the two words were also similar in that classification; a hit was recorded if the two words appeared there under the same topic number.

Following the analysis, the syntactic method was found to give significantly better results for the 600 most common nouns in the corpus, while the window-based approach was superior for the remaining words. It was therefore concluded that no single statistical technique is suited to the analysis of words from all ranges of frequencies in a corpus. The argument was made that, for frequent words, the sort of fine grained analysis permitted using the syntactic method provided enough information to enable similarity to be judged. For less frequent words, on the other hand, the window based method began to provide more information, although of a less exact nature.

Given the attractiveness of Roget's Thesaurus as a basis for evaluating the success of analyses based on different sets of parameters or different techniques, it was decided that a method of the same general type as that of Grefenstette should be adopted in attempting an evaluation of the analyses presented in the last chapter.

5.3 Details of the Approach

In deciding how best to use Roget's Thesaurus[10] for evaluating distributional analyses of corpora, it initially seemed desirable to do this using the clusters revealed in the

[10] In carrying out the evaluation, the Project Gutenberg Etext version of the 1911 edition of Roget's Thesaurus was used. This contains over 1000 words added to the 1911 edition, although many modern words are still absent from it. In total, it contains more than 30,000 unique words allocated between 1000 categories.

dendrograms. That is, it would be attractive to select a particular target word from a dendrogram and then its nearest neighbour or neighbours.

However, given the type of cluster analysis performed in these analyses, this is not straightforward to achieve. In general, a word is not added to a cluster on the basis of its similarity to a single word, but its similarity to an average calculated over a number of words in an existing cluster. This means that the position of a word which appears closest to a particular target word in a dendrogram is not necessarily determined by its proximity to the target word alone, but is also influenced by its proximity to other words already placed in the dendrogram.

In the light of this, it seemed appropriate to adopt a more straightforward approach, in which the nearest neighbours to target words really would be considered. For the most frequent 1000 words in the corpus, the 10 nearest neighbours were calculated. It seemed desirable to use this information, rather than that contained in the dendrograms, to conduct an evaluation using Roget's Thesaurus. As noted above, Grefenstette (1993) also approached the problem of evaluation in this manner, using nearest neighbours.

Rather than restricting the evaluation to the 50 randomly selected words presented in the previous chapter (which were selected to make presentation of the data manageable), there seemed no reason not to conduct the evaluation over all of the 1000 most frequent words in the corpus. Furthermore, rather than considering only the nearest neighbour to each target word (as with Grefenstette (1993)), the 10 nearest neighbours were taken into account. This decision was made because considering only the first nearest neighbour may mean that the evaluation may not give justice to the success of the classification method used, since inspection suggests that more words than just the first nearest neighbour are closely related to the target word.

For each of the most frequent 1000 words in the corpus, then, the 10 nearest neighbours were included in the evaluation. As with Grefenstette's (1993) method, a hit would be recorded on each occasion when one of these nearest neighbours

occurred under the same topic number (from the 1000 topic numbers used by Roget's Thesaurus) as the target word in the thesaurus. For each target word, the maximum score that could be achieved would be 10. This gives a theoretical maximum score for the analysis as a whole of 10000, the situation in which each of the 1000 target words considered achieves a hit for each of its 10 neighbours. In fact, no analysis could achieve such a score because some target words do not appear in the thesaurus and could not therefore achieve a hit for any of their neighbours. This situation arises with closed class words in particular, as discussed further below. The actual maximum score which could be achieved with the particular set of words used was not calculated, since the evaluation is intended primarily as a fairly crude means with which to compare the appropriateness of the classifications achieved by the various analyses. In other words, it is the *relative* scores achieved which are of most interest here.

The classification of words in Roget's Thesaurus, unlike the classification being evaluated here, permits a target word to appear *more than once*. This allows various senses of the target words to be represented, whereas in the analyses being evaluated each target word appears only once, and the representation for each of these 'smears together' all the senses into which it might be decomposed. This is certainly an undesirable feature of the present method and is one which is discussed in more detail in Chapter 6.

In an attempt to adapt Roget's Thesarus to this sort of classification, and thus to make the evaluation workable, the categories in the thesaurus which contained each target word were pooled into one large category. This large category would then contain words related to all the senses of the target word recognized by the thesaurus, and would hopefully correspond roughly to the collection of senses incorporated into the representation of the target word used in our analyses. The procedure was then simply to count how many of the 10 nearest neighbours of the target word also appeared in the pooled category from Roget's Thesaurus. Where a neighbour appeared more than once in this pooled category, the score was still incremented by only one.

In table 5.1 below, the pooled category for the target word 'kind' is presented for the purposes of illustration. Roget's Thesaurus classifies this target word into two categories, corresponding to two discrete senses of the word. Since the analyses being evaluated would not distinguish between these, they would be pooled for the purposes of evaluation and considered as a single category.

Table 5.1: Example of Pooled Categories from Roget's Thesaurus

Section Number and Heading from Roget's Thesaurus	Words Contained Within Category
75. *Class*	class, division, category, categorema, obs3, head, order, section, department, subdepartment, province, domain, KIND, sort, genus, species, variety, family, order, kingdom, race, tribe, caste, sept, clan, breed, type, subtype, kit, sect, set, subset, assortment, feather, kidney, suit, range, gender, sex, kin, manner, description, denomination, designation, rubric, character, stamp predicament, indication, particularization, selection, specification, similarity.
906. *Benevolence*	benevolence, Christian charity, God's love, God's grace, good will, philanthropy, 910, unselfishness, 942, good nature, good feeling, good wishes, kindness, kindliness, loving-kindness, benignity, brotherly love, charity, humanity, fellow- feeling, sympathy: goodness of heart, warmth of heart, bonhomie, kind- heartedness, amiability, milk of human kindness, tenderness, love, 897, friendship, 888, toleration, consideration, generosity, mercy, (pity), 914, charitableness, bounty, almsgiving, good works, beneficence, "the luxury of doing good " , Goldsmith, acts of kindness, a good turn, good offices, kind offices good treatment, kind treatment, good Samaritan, sympathizer, bon enfant, Fr, altruist, be benevolent, have one's heart in the right place, bear good will, wish well, wish Godspeed, view with an eye of favor, regard with an eye of favor, take in good part, take an interest in, feel an interest in, be interested in, feel interested in, sympathize with, empathize with, feel for, fraternize, (be friendly), 888, enter into the feelings of others, do as you would be done by, meet halfway, treat well, give comfort, smooth the bed of death, do good, do a good turn, benefit, (goodness), 648, render a service, be of use, aid, 707, benevolent, KIND, kindly, well-meaning, amiable, obliging, accommodating, indulgent, gracious, complacent, good-humored, warm-hearted, kind-hearted, tender-hearted, large-hearted, broad- hearted, merciful, 914, charitable, beneficent, humane, benignant, bounteous, bountiful, good-natured, well-natured, spleenless, obs3, sympathizing, sympathetic, complaisant, (courteous), 894, well-meant, well-intentioned, fatherly, motherly, brotherly, sisterly, paternal, maternal, fraternal, sororal, obs3, friendly, 888, with a good intention, with the best intentions, Int. Godspeed! much good may it do! ," act a charity sometimes " , Lamb, " a tender heart, a will inflexible " , Longfellow, de mortuis nil nisi bonum , Lat: say only good things about the dead, don't speak ill of the dead, " kind words are more than coronets " , Tennyson, quando amigo pide no hay manana, Lat, " the social smile, the sympathetic tear " , Gray.

Sections in Roget's Thesaurus containing target words and neighbours were only used if the target word appeared in the section as a single word rather than as part of a phrase. Thus, section 75 (*class*) was selected as one of the classes containing the word 'kind' above, but section 765 (*request*) was not because the word only occurs there as part of the phrase "would you be so kind as to".

In guarding against the inclusion of inappropriate occurrences of the target words, we also prevent the inclusion of many occurrences of closed class words (which often form part of multi-word entries) in the thesaurus. This would seem to be a desirable state of affairs, since it is the meaning of the phrase as a whole, rather than the

function words within it, which is of importance in the thesaurus. Furthermore, the meanings of closed class words tend to be of a much more nebulous nature than for other types of words, and they often cannot justifiably be placed in a particular category or categories on the basis of meaning. Indeed, no attempt is made in Roget's Thesaurus to classify many of the closed class words, and their occurrence is largely restricted to multi-word entries.

This difficulty in establishing the meaning of closed class words is exemplified experimentally in the work of Jones (1985), in which subjects were asked to rate various words on the basis of 'Ease of Predication', a concept which was operationalized as 'ease of putting words into simple factual statements'. It was found that the Ease of Predication scores for function words were significantly lower than for all other types of words considered (high-imagery nouns, low-imagery nouns, adjectives, and verbs)[11].

The sections in Roget's Thesaurus containing each of the 1000 target words were identified. After being combined into larger sections, these were then searched for the occurrence of the 10 neighbours in each case. The whole evaluation procedure was implemented as a program on a UNIX computer, and is summarized in table 5.2 below.

[11]Within the group of function words considered by Jones, personal pronouns were found to have significantly higher Ease of Predication scores than the other types of function words used (relative pronouns and interrogatives, prepositions and conjunctions, and auxiliaries).

Table 5.2: Procedure Used for Matching Nearest Neighbours With Categories in Roget's Thesaurus

Stage 1:
For each target word considered,
1. Identify the sections in Roget's Thesaurus which contain the word as a single entry.
2. Combine these into a single section containing the entries from each of the sections identified in the previous step.
Stage 2:
For each target word considered,
3. Locate the combined section created in step 2 above.
4. Increment the score by 1 for each of the target word's neighbours which occurs in the section.

5.4 Results

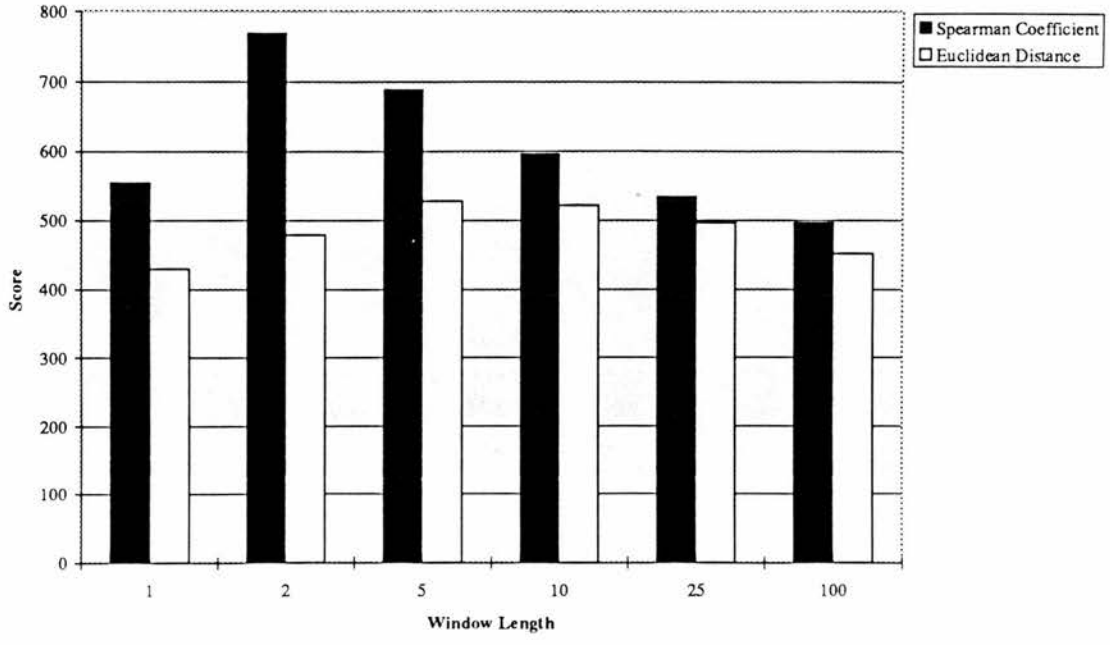
The results of the evaluation are presented in table 5.3 below

Table 5.3: Results of Evaluation Using Roget's Thesaurus as a Benchmark

Method of analysis			
Analysis Number (Chapter 4)	Distance Metric	Window Length	Score
1	Spearman Rank Correlation Coefficient	1	554
2	Euclidean Distance	1	429
3	Spearman Rank Correlation Coefficient	2	769
4	Euclidean Distance	2	478
5	Spearman Rank Correlation Coefficient	5	689
6	Euclidean Distance	5	527
7	Spearman Rank Correlation Coefficient	10	595
8	Euclidean Distance	10	521
9	Spearman Rank Correlation Coefficient	25	534
10	Euclidean Distance	25	496
11	Spearman Rank Correlation Coefficient	100	496
12	Euclidean Distance	100	451

These results are also presented in figure 5.1 below.

Figure 5.1: Results of Evaluation Using Roget's Thesaurus as a Benchmark



Before examining whether these differences are significant ones, it is important to make a check on the validity of the clusters resulting from each of the analyses. In each of the above analyses, the 10 nearest neighbours for each of 1000 target words were considered. We need to ensure that these neighbours are reasonably different for each of the target words; in general, we would not expect pairs of target words to have very similar neighbours. In the worst case, if the clustering procedure had been seriously flawed, all target words might have been assigned the same set of nearest neighbours.

To ascertain the extent of the overlap between the clusters in each analysis, the following check was carried out. The 1000 target words used in each analysis were divided into two groups of equal size, with the words in each group being matched for frequency in the corpus. For each of the target words in each group, those neighbours which had contributed to the score for the analysis (that is, those which had appeared in the relevant section of Roget's Thesaurus) were also listed. The groups were then examined for 'shared neighbours' by calculating the number of times a target word in one group shared a neighbour with its frequency-matched counterpart in the other group. In general, we should expect this number to be low; there is no reason why

two frequency-matched target words should be particularly similar in terms of the contexts in which they occur. The results of this check are shown in table 5.4. In this table, the total number of neighbours for the 500 target words in each of the two groups is given, along with the number of times that a neighbour was shared between two frequency-matched target words.

Table 5.4: Number of Neighbours Shared by Randomly Chosen, Frequency-Matched Target Words

Analysis Number (Chapter 4)	Number of scoring neighbours in first group	Number of scoring neighbours in second group	Number of 'shared scoring neighbours'
1	273	281	0
2	212	217	0
3	370	399	1
4	241	237	0
5	338	351	0
6	267	260	1
7	298	297	1
8	265	256	2
9	267	267	1
10	240	256	1
11	248	248	1
12	216	235	0

As table 5.4 makes clear, a very small proportion of the neighbours which had contributed to the score in each analysis was shared between the pairs of target words. This indicates that the neighbours contributing to the score in each case were indeed quite different for each target word. Having established this, we can now proceed to examine the results indicated in figure 5.1 in more detail.

The skewed nature of the data within each of the conditions shown in figure 5.1 (skewed towards low scores) indicated that it would be most appropriate to examine the differences here using non-parametric statistical procedures.

Using the Friedman non-parametric analysis of variance, a significant overall effect was found for the distance metric used, with the Spearman metric giving better scores ($\chi^2 = 39.0427$, $p<0.01$). The differences between the individual analyses were further assessed using the Wilcoxon signed ranks test, with an appropriate correction for

carrying out multiple comparisons (Siegel and Castellan, 1988). The Wilcoxon statistic resulting from the test was converted to a z-score in each case.

Using the Euclidean distance metric, a window length of 5 words either side of the target word gave a significantly better score than a window length of 1 word ($z=-3.6665$, $p<0.05$, 2-tailed). Similarly, a window length of 10 scored more highly than a window length of 1 word ($z=-3.5815$, $p<0.05$, 2-tailed). All other comparisons between analyses using the Euclidean distance metric were non-significant.

Using the Spearman correlation coefficient as the distance metric, a window length of 2 words was found to give a significantly better score than a window length of 1 word ($z=-7.3851$, $p<0.05$, 2-tailed), 10 words ($z=-6.3267$, $p<0.05$, 2-tailed), 25 words ($z=-7.6313$, $p<0.05$, 2-tailed), and 100 words ($z=-8.1905$, $p<0.05$, 2-tailed). The use of a window length of 2 words did not give a significantly better score than a window length of 5 words.

5.5 Discussion

Using Roget's Thesaurus as a benchmark and the procedures described earlier for evaluating each of the analyses carried out in Chapter 4, two main findings have emerged. Firstly, use of the Spearman Rank Correlation Coefficient as a distance metric appears to give rise to a categorization of the target words in the corpus which is in closer accordance with Roget's Thesaurus than does use of the Euclidean distance metric. Secondly, a window length of 2 or 5 words either side of the target word results in a categorization which is closer to that contained within Roget's Thesaurus than do window lengths of other sizes.

As we noted earlier, the evaluation carried out is best regarded as one which is relative, rather than absolute. The scores obtained were small when compared to the theoretical maximum, and this is likely to be due in part to the decision to consider the 10 nearest neighbours for each target word[12]. If this number were reduced, the ratio

[12] An experiment is in progress to compare human subjects' evaluations of these analyses with those produced by the method using Roget's Thesaurus.

between the actual score obtained and the theoretical maximum could conceivably be reduced. In any case, the task of producing 10 neighbours for each target word which accord with Roget's Thesaurus may be a very difficult one, even for human beings. Despite the fact that the scores were small relative to the theoretical maximum, they are large relative to what would be expected on the basis of chance. Taking a simplified example to illustrate this, consider choosing 10 nearest neighbours for a target word at random; the probability[13] that just one of these neighbours coincides with a previously determined word which appears in the relevant section of Roget's Thesaurus (and which is, of course, also present in the set of target words) is only 3.9×10^{-21}. Further reasons for the low performance relative to the theoretical maximum are likely to be that not all the target words were in fact present in Roget's Thesaurus, and that the set of semantically related words available to be placed in the categories of the Thesaurus would generally have been larger than the restricted set of 1000 words used here. At the same time, it is possible that some of the Roget's categories may have contained *fewer* than 10 words, again making a maximum score impossible to achieve.

Taking the issue of the distance metric first, it should perhaps be noted that the Spearman metric was initially chosen for use after examination of the work of Finch and Chater (1992a, 1992b), which indicated that this would be a useful measure to employ. At first sight, it is perhaps surprising that it performed more successfully than the Euclidean measure, because the Spearman metric uses less information about the vectors being compared than does the Euclidean one. With the Spearman metric, which does not pay attention to the absolute magnitudes of the vector components, it is only the direction of the vectors concerned which is taken into account. On the other hand, with the Euclidean metric, which does pay attention to the magnitudes of the vector components, the lengths as well as the direction of the vectors are considered. The Spearman metric, which involves the ranking of vector components rather than taking their absolute magnitudes, is consequently likely to be less sensitive

---

[13] This probability is $^{10}C_1 \left(\frac{1}{999} \times \left(\frac{9}{999} \right)^9 \right)$ since, for each target word, there are 999 possible neighbours to choose.

to noise than is the Euclidean one. Whilst the vectors were normalized for the frequency of the target word they represent (which would otherwise be a source of noise), it may be that noise was nonetheless present amongst the co-occurrence probabilities for the context words. As we noted in Chapter 4, recording these probabilities for as many as 1000 context words may introduce more unreliable statistics than if a smaller number were used.

An important general conclusion here, then, is that the choice of distance metric in analyses of the type we have carried out is not immaterial. When evaluated against the benchmark of Roget's Thesaurus, the Spearman metric is clearly superior to the Euclidean one. The clear differences between the results obtained using the two metrics are apparent when inspecting the relevant tables of nearest neighbours presented in Chapter 4. It is often the case that the set of 10 nearest neighbours presented for a target word when using the Spearman metric will be considerably different from the set presented for the same word under an equivalent analysis using the Euclidean measure, even though both sets will exhibit a degree of semantic relatedness to the target word. Future work in this area would be well advised to bear in mind that the results obtained are dependent on the distance measure used to produce them, and more exhaustive future investigations into the influence of particular distance metrics on results would be extremely valuable.

Having made these observations, the Spearman metric is nonetheless a relatively straightforward measure which does not make sophisticated assumptions about the data being analysed. Its use in the analyses presented in Chapter 4 has revealed rich structure amongst the vectors in the high-dimensional space employed to explore the informativeness of statistical context. As a result, it seems reasonable to suggest that it is a useful metric to use in analyses such as those that have been presented here.

The Spearman metric was found to give rise to categorizations closer to those found in Roget's Thesaurus than the Euclidean metric. Within the analyses carried out using the Spearman metric, those that used a window length of between 2 and 5 words either side of the target word were found to provide the 'best' categorizations. This finding is of interest from a psychological perspective, since it suggests that, in natural

language, the amount of context around each word to which attention must be paid has an optimal size. If this size is expressed in terms of the number of words, as it has been here, it appears to be the relatively small distance of between 2 and 5 words.

Gale, Church and Yarowsky (1992) have noted that a 5 word context can be justified on the basis that human subjects can carry out word-sense disambiguation using a context of this sort of size, but have not offered any direct empirical support for this.

The optimal amount of context indicated in the analyses presented in this chapter does suggest some similarity with a particular aspect of the work of Baddeley (1990). Baddeley has proposed a psychological model of human memory, in which a 'phonological loop' plays an important part as a subsystem. This loop contains a short-term memory store which can retain speech-based information for up to 2 seconds, after which the memory trace fades and becomes unretrievable unless rehearsal is carried out to refresh it. Baddeley psychological evidence which suggests that the limit of 2 seconds is a fundamental characteristic of human short-term memory capability. For example, he reports that as the number of syllables in a list of words increases, the probability that the words will be recalled correctly by subjects decreases; in other words, as the spoken duration of the words increases, fewer of them can be stored before the 2 second limit has been reached. If there is any possible connection between Baddeley's 2 second limit and the optimum window size identified here, we need first to ascertain approximately how many English words would typically be uttered in 2 seconds.

Lenneberg (1967) was concerned with the issue of how fast English speakers can speak, and reviewed various studies seeking to answer this question. Following analysis of three different newscasters, Lenneberg concluded that their rate of speech was approximately 14 phonemes per second, or about 5.9 syllables per second. We now need to know how many words per second this corresponds to. Shillcock, Hicks, Cairns, Levy, and Chater (1995) have recently conducted an extensive empirical study of the statistics of the English language, by means of an analysis of the London-Lund corpus. Following the analysis, these authors reported that the average number of

syllables per English word was approximately 1.33. On the basis of these studies, then, the average rate of speech of an English speaker may be taken as roughly 4.4 words per second. Within Baddeley's 2 second limit, we could therefore expect about 8.8 words to have been uttered. If we remember that our window length of 2 to 5 words corresponds to a length of 4 to 10 words when context on both sides of the target word is considered, this value of 8.8 words does fall within the optimum size we have established.

It could be that the length of Baddeley's phonological loop is not in fact related to the amount of context used for distributional work of the type discussed here, and that no similarities should therefore be expected between the two approaches. However, it is worth noting that Baddeley has argued that the loop may be an important aspect both in the comprehension of sentences, and in acquiring a vocabulary. The analyses presented in this chapter are, of course, very much concerned with the possibility of acquiring particular aspects of a vocabulary from exposure to spoken language, and so it is of interest to enquire whether Baddeley's psychological findings really can be related to a basic statistical constraint of the English language. If they can, the close correspondance between the optimum window length identified here and Baddeley's limit may be no coincidence. We must bear in mind, nonetheless, that the assessment criterion provided by Roget's Thesaurus is more of a relative one than an absolute one, and its precision is open to question. We must also remember that, although distributional language research has tended to incorporate it since it does provide useful information about the behaviour of the target word, the use of right-sided context is debatable on the grounds of psychological plausibility. Some research does suggest, however, that it is of importance; Bard, Shillcock, and Altmann (1988), for example, have noted that in the 'gating' paradigm, around 20% of words could only be identified by subjects in the presence of following words.

Whether or not the optimal amount of context required for 'good' conceptual structure is related to the length of speech which the human memory system is capable of storing will perhaps not become clear for some time. However, research on this

issue would seem to be worthwhile. Baddeley does himself suggest that his 2-second limit may ultimately have derived from human exposure to spoken language:

"... reading surely developed too recently for this to offer a plausible explanation as to why an articulatory loop system should have evolved. A more plausible explanation might be to suggest that the phonological loop developed in the process of the evolution of speech production and comprehension (Baddeley, 1990; p88)"

5.6 Conclusions

In this chapter we have presented an objective assessment of the analyses presented in Chapter 4, using Roget's Thesaurus as a benchmark. The results indicated the superiority of the Spearman Rank Correlation Coefficient over the Euclidean distance metric, and that a window length of between 2 and 5 words either side of the target word produced a closer match with Roget's Thesaurus than other window lengths. These findings suggest that a relatively short window length may be optimal when using statistical information to learn about the relationships between word meanings, and are in close agreement with Baddeley's (1990) description of the 'phonological loop'.

Following the encouraging results obtained in this chapter and the last, in the next chapter we shall reassess the 'standard' statistical approach, consider its limitations, and consider the possibility of improvements to this approach of categorizing words on the basis of meaning.

6. *AN ALTERNATIVE TO THE 'STANDARD' APPROACH?*

6.1 A Fundamental Problem with the 'Standard' Analyses

Towards the beginning of this thesis, the aim was set out of attempting to explore the extent to which a categorization of word meanings could be achieved using simple statistical methods which use intralinguistic information alone. In Chapter 4, analyses of this type were explored in detail using a range of different parameters. The results of the analyses were presented both in the form of dendrograms and in the form of lists of nearest neighbours for the various target words under consideration. It was noted in Chapter 5 that a more objective method of assessing the reasonableness of the categorizations obtained would be desirable, and a method of comparing the groups of nearest neighbours with the groupings contained within Roget's Thesaurus was presented. The results suggested that a relatively short context window of 2-5 words would produce the 'best' results.

Whilst the approach of comparing the categorizations obtained with those in Roget's Thesaurus is undoubtedly useful as an indication of relative performance, the results of such analyses (whose approach has been popular with numerous researchers) are unlikely ever to be very much more successful until a fundamental problem has been resolved. This problem, as Huckle (1995) has noted[14], is that each target word in the analysis is permitted only a single representation, which is a vector containing probabilities calculated over all occurrences of the target word in the corpus. As such, the vector representation is some kind of 'average' or 'smearing' over any separate senses into which the target words may have fallen. This difficulty applies also to syntactic categorization, and Schütze (1993a) makes it explicit:

> "Ambiguity is a problem ... because the two components of an ambiguous vector can add up in a way that makes it by chance similar to an unambiguous word of a different syntactic category (p254)"

We know, however, that polysemy is an important feature of English and other natural languages. In many cases, therefore, it may be inappropriate to force only a single representation for each word being considered. To take a commonplace example provided by Gallant (1991), the word 'star' could refer to a celestial body, a

[14] A copy of this paper is included in Appendix D.

Hollywood personality, or the act of writing an asterisk. Not only are these different senses of the word different in meaning, but they also differ with respect to syntactic category. If we would like the word 'star' to be categorized sensibly with other words of similar meaning on the basis of distributional statistics, it would seem important to stop combining separate senses of the word into a single representation. Despite this, the use of a single representation for each word has continued to be a common approach amongst those working in this field.

The use of a single representation for each target word can perhaps be defended initially on the basis that it would be as well to assess the appropriateness of the context vector approach with a simplified methodology at first in order to determine whether such an approach is likely to work at all. We have seen in Chapters 4 and 5 that the use of even very simple distributional statistics can indeed allow surprisingly rich semantic categorizations to be obtained from natural language corpora. It now seems appropriate to move on further to a stage in which the restriction of having only a single representation for each target word is no longer enforced. This undoubtedly makes the whole problem much more complex, but at once much more like that which would be faced by a language learner exposed to a large amount of natural language data. Of course, it must be conceded that native speakers of a language will have more than just the intralinguistic information to assist them in the task of disambiguation (see, for example, Gerrig and Littman (1990)), but for present purposes we shall restrict our attention, as before, to the single source of information provided by the language structure itself.

In Chapter 7, this problem is addressed using an unsupervised neural network implementation. However, we shall first consider some previous approaches towards word-sense disambiguation. It is important to note beforehand, however, the existence of two slightly different settings in which word-sense disambiguation is an issue to be confronted. The first is a situation which has already been alluded to; it is the case in which the particular concept or meaning of a word must be determined, given its context of occurrence. Thus, we might be faced with the problem of deciding the meaning of the word 'star' in the utterance 'the astronomer married the star'. The

second situation is one in which we need to determine the correct lexical item to be used when translating a particular lexical item in one language into a second language. In such a case, the second language would distinguish between two (or more) lexical items where the first would not. Thus, when translating from English to French, we may have to decide whether 'I know him well' should be translated as 'je le sais bien' or as 'je le connais bien'. Whilst these two situations are superficially different, the problem being dealt with is essentially the same and involves a one-to-many mapping; given a word token in a particular context we need to decide which of a number of possible alternatives is the most appropriate one with which to represent that word.

6.2 Previous Work with Word-sense Disambiguation

Gallant (1991) suggested that word-sense disambiguation could be carried out using context vectors on the basis of syntax, context, common usages of words, and world knowledge. He felt that the use of such an approach could easily be implemented with existing natural language systems and would enable machine translation systems which do not currently consider context to enjoy increased accuracy. Gallant proposed that a vector be constructed for every individual meaning of polysemous words, and that the vector for the word could be considered as the sum of the vectors for these different meanings. To choose the appropriate sense for a particular occurrence of a word, Gallant proposed that this could simply be achieved by choosing the meaning having the vector closest (measured using the dot product) to the word's context vector in the text. Assuming that the vectors are normalized, this would, as Gallant notes, correspond to choosing the meaning that is closest in direction to the word's context vector. Gallant outlined a neural network implementation which could potentially be used to perform this task, and the neural network developed and tested in Chapter 7 is similar to this.

Gallant (1991) considered that, in addition to the benefits for machine translation systems, a context vector approach to word-sense disambiguation would also be of interest to those concerned with psychological modelling. As he has pointed out,

"A number of experiments are suggestive of feature-based representations because humans make simple and unexpected errors that would be explained by the use of such representations at an early level of cognitive processing. The basic idea is that a simple feature-based representation does not keep track of *relations among the features* and this would cause certain types of errors (p305-6)."

Whilst Gallant's suggestions are an interesting indication of the growing interest in corpus-based statistical context vector approaches at the beginning of the 1990's, they are of course only proposals which were intended to stimulate research in this area. Since his proposals were made, such research has made rapid progress.

From the point of view of language acquisition, which has been one of the main themes of this thesis, word-sense disambiguation is also an issue of fundamental importance, since it is a problem with which language learners are perpetually confronted. Writing at much the same time as Gallant, Sonaiya (1991) has stressed this from an Applied Linguistics perspective, considering the difficulties presented during the learning of a second language. Sonaiya's model of second language acquisition has close parallels with the approach taken in this thesis, and the main points of this are worth considering in detail. Sonaiya has pointed out that the errors made by the speaker of one language when learning another have traditionally been explained in terms of errors of transfer which are due to interference from the native language. However, Sonaiya makes the suggestion that it might be more appropriate to consider such errors from a semantic point of view:

> "This approach views words in terms of the semantic relations that exist among them. Thus the word that has been erroneously used by the learner is analyzed in terms of the relationship it bears to the item that should have been used ... (p274)"

Sonaiya proposes a Continuous Lexical Disambiguation Model which seeks to capture the continuous process of refining and readjusting of the boundaries between lexical items, and points out that the conventional view which assumes that lexical items are learned by rote is unlikely to be appropriate. The reason for this is that no pair of languages has yet been discovered in which there is a strict one-to-one mapping from names of concepts in one language to names of those concepts in the other. The errors made by adults in learning a second language often reflect the difficulties this presents; when a particular concept is lexicalized in a different manner in the second language to the way it is lexicalized in the native language, there can be problems in deciding which lexical item to use for that concept when speaking the new language. Sonaiya notes that with English speakers learning Spanish, for example, difficulties can arise because the verb 'to be' maps to two verbs in Spanish: 'ser' and 'estar'. Given the ubiquity of this kind of situation, Sonaiya stresses that

vocabulary acquisition must crucially be regarded as the learning of the lexical relationships between semantically related items in the language being learned, and even proposes that the semantic fields into which words could be placed might be represented as a multidimensional space[15]. This form of representation is of course of a type which has been explored throughout this thesis.

Sonaiya has also noted that when making errors in selecting a lexical item in a second language, language learners are often well aware of the correct lexical item. The difficulty for them appears to lie in distinguishing the different instances in which the two (or more) words should be used. In this thesis, these different instances are described in terms of the contexts in which words can occur and they form the basis for the representational system used here. Sonaiya proposes that as knowledge increases, knowledge about the contexts in which words occur permits a structural reorganization of the vocabulary of the target language (as it is understood by that individual) to take place. Much the same process is also proposed for the acquisition of a first language:

"... the development of a child who starts by referring to all four-legged animals as "dog" but who later acquires the distinctions that exist between a dog and other animals can be represented by the model presented here. The first stage will be a situation in which the item *dog* is the only one that occupies the space belonging to four-legged animals. As the distinction between a dog and other animals is acquired, boundaries are set up and continuously readjusted to reflect the current stage of the knowledge of the child (p281)."

In this way, the process of acquiring a second language or, indeed, a first language, is seen as a process of continuous lexical disambiguation. Sonaiya suggests that exercises on lexical disambiguation might, on the basis of this sort of conceptualization of language learning, form a useful part of vocabulary teaching.

From a more computational perspective, Brown, Della Pietra, Della Pietra, and Mercer (1991) have considered a statistical approach to word-sense disambiguation. In confronting the problem of translating French sentences into English, they used a Bayesian approach in which an 'alignment' was first constructed between possible English and French sentences using the Canadian parliament's Hansard publication as

[15] The dimensions used to represent words in this multidimensional space, however, are not of the statistical type used here. Instead, Sonaiya proposes the selection of certain dimensions such as 'register', 'animacy', 'concreteness', or 'abstractness'.

a corpus. This was required in order to calculate the most likely French translation, given a particular English sentence (which is a statistic required by the Bayesian approach they used). In the alignments used (which compared an English and a French sentence in each case), each French word, or group of French words was 'connected' by a line to an English word. From these alignments, the probability could be calculated that a particular English word would be connected to a particular French word and the mutual information beween the two words could then be computed. The method then proceeded by labelling a word with a sense which served to increase the mutual information between the members of a connection. Promising results were obtained in translating between French and English using this type of information-theoretic approach.

Similar work was also conducted by Brown, Lai, and Mercer (1991). 'Parallel' French and English corpora were used in an attempt to extract pairs of sentences which were translations of each other. The texts were again taken from the Hansard proceedings of the Canadian parliament. The problem of alignment was tackled by first aligning certain anchors between the two texts. These anchors made use of comments in the text which identified the person speaking or the time at which the utterance was made. Once these anchors had been lined up between the two corpora, the individual sentences between the anchors had to be aligned. Not only were Brown et al. confronted with the lack of a one-to-one mapping between the words in the corpora, but also between the *sentences*:

"since the number of sentences in the French corpus differs from the number in the English corpus, it is clear that they cannot be in one-to-one correspondence throughout (p172)".

It was observed, however, that in sentences which were translations of each other, the number of word tokens in each was correlated; thus a longer sentence in French tends to translate to a longer sentence in English, while a shorter sentence in French would be likely to have a correspondingly shorter English sentence as its translation. This information was then employed in estimating the appropriate alignments to be made between the sentences in the two corpora. Several million sentences were aligned, and when 1000 sentence pairs were subsequently checked by hand, an accuracy of around 99% was recorded.

Gale and Church (1991) also used the Canadian Hansards in investigating a method for aligning French and English sentences. Again, the correlation in sentence length between the two languages was exploited, although in this case length was measured in terms of the number of *characters*, rather than the number of words as with Brown, Lai, and Mercer (1991). Gale and Church's method was initially tested on a trilingual corpus in English, French, and German, taken from economic reports issued by the Union Bank of Switzerland. On this corpus, only about 4% of the sentences were aligned incorrectly. The number of errors, furthermore, was found to be roughly the same when translating from English into French or to German, indicating that this type of approach to alignment may be relatively independent of the particular languages involved.

In reviewing work within Computational Linguistics on word-sense disambiguation, Gale, Church, and Yarowsky (1992) have commented that such efforts show promise:

"Much of this work offers the prospect that a disambiguation system might be able to input unrestricted text and tag each word with the most likely sense with fairly reasonable accuracy and efficiency, just as part of speech taggers ... can now input unrestricted text and assign each word with the most likely part of speech with fairly reasonable accuracy and efficiency (p249)".

They also note that parallel texts, like those already described, can be useful in a slightly different way than we have already encountered for word-sense disambiguation. Rather than using the two texts as a basis for learning to translate appropriately from one language to the next, the fact that a word in one language may translate into more than one word in the second can itself be used as a useful indicator of the different senses into which that word can be divided in the first language. Thus, the two approaches to word-sense disambiguation (deciding upon the intended sense of a word given its context within a single language, or translating from one language to another) could potentially be combined to allow information from the second language to suggest which sense of a word is intended in the first. Gale, Church and Yarowsky do caution, however, that this approach may not be ideal because

"... the assumption that differences in translation correspond to differences in word-sense has always been somewhat suspect (p251)".

To circumvent the possible disadvantages of using the Canadian Hansards for word-sense disambiguation, Yarowsky (1992) used Roget's Thesaurus as the source of

information for his approach to the problem. The central idea here was to select the category in the thesaurus which would be most likely given the context in which a word appears. It will be recalled that a related approach was used in Chapter 5 in evaluating the performance of vector-based representations for words in the Wall Street Journal. However, in that evaluation each word had only a single representation; Yarowsky aimed to allow each word token to have different representations depending upon the context of use.

Yarowsky trained his system on 10 million words of Grolier's Encyclopedia. For every occurrence in the corpus of each word in a particular category of Roget's Thesaurus, the 100 surrounding words were recorded. Once this information had been obtained, 'salient' words could be identified. These were defined as words which appeared significantly more often in the context of a category than at other points in the corpus, and which were felt therefore to be relatively good indicators of the category concerned. For example, the salient words for the 'tools/machinery' category of Roget's Thesaurus were found to include the words 'tool', 'machine', 'blade', 'device', 'pump', and 'tooth'. The occurrence of salient words in the neighbourhood of an ambiguous word in the corpus was then used to weight the probability of it belonging to a particular category in Roget's Thesaurus. This method was found to give very encouraging results, with the senses of selected polysemous words in the corpus being correctly determined on the majority of occasions.

Schütze (1992) has also proposed a method for peforming word-sense disambiguation using statistical vector representations for words derived from large corpora. Schütze has noted that words represented by vectors containing similar components will be relatively more highly correlated than words represented by vectors containing dissimilar components, and that words which are highly correlated in this manner can be regarded as being similar to one another. These sorts of assumptions are of course familiar ones that are central to the statistical vector approach discussed in Chapter 4. The correlation measure proposed by Schütze is the cosine of the angle between the vectors, and Schütze notes that:

"similarity between vectors has then a straightforward visual equivalent: Closeness in the multidimensional space ... (p101)".

Rather than using this type of approach to distinguish words, however, Schütze proposes its use in distinguishing between senses of words. Given the context of a word at a particular point in the corpus, the problem is to decide which sense of the word is intended. If the possible senses of the word are known, it may also be possible to say which words would be likely to occur in the context vector for each of those senses. To use Schütze's example, we may be interested in disambiguating a particular occurrence of the word 'interest'. If we know that the word 'interest' has two possible senses which are PERCENT (denoting the sense of "charge on borrowed money") and CONCERN ("a feeling that accompanies or causes special attention"), we may also know which words are likely to occur in the context vector for each of these senses. Schütze asserts that 'soar' is more likely to occur in the vector of the PERCENT sense, and 'sport' is more likely to occur in the vector of the CONCERN sense. The context vector for 'interest' can then be compared with 'soar' and 'sport' dimensions of the space in which the vector is located, and the one to which it is most closest will define the sense being used. In practice, co-occurrence with more than just two words would need to be considered.

In using this type of approach, Schütze used a singular value decomposition procedure to reduce the dimensionality of his vectors. This enabled him to overcome the problems of the data being noisy and the large amount of space which would otherwise have been required for storing it. Empirical work showed that the overall approach worked very well in disambiguating occurrences of a number of selected words in a corpus taken from the New York Times.

Schütze (1993b) followed this with a further investigation of word-sense disambiguation using vector representations for target words. He represented the words in terms of their co-occurrences with 'letter fourgrams' (see Chapter 7 for further discussion of this method), rather than with other words. His approach was then the fairly primitive one of clustering all the individual context vectors for particular target words in the corpus, rather than combining them to form a single vector in the usual way. In order to ascertain the nature of the resulting clusters, 10 to 20 members of each were inspected. The sense of a particular word could then be

disambiguated by assigning it the sense of the cluster nearest to its context vector. It was found that for particular target words occurring in the corpus, the resulting clusters did indeed reflect particular senses in which the words could be used. The word 'capital', for example, was clustered into groups labelled as 'goods' and 'seat of government', and 'space' was clustered into groups labelled as 'area, volume' and 'outer space'. Word-sense disambiguation using these senses was found to be highly accurate in general.

Schütze (1995) carried out further work on ambiguity, this time considering the problem of *syntactic* ambiguity. As he points out, the majority of approaches to linguistic classification are similar in that they

"... classify *words* instead of individual occurrences. Given the widespread part-of-speech ambiguity of words this is problematic. How should a word like "plant" be categorized if it has uses both as a verb and as a noun? How can a categorization be considered meaningful if the infinitive marker "to" is not distinguished from the homophonous preposition? (p141)"

However, Schütze (1995) has also noted that in many cases in which disambiguation is to be carried out, information about the word to be disambiguated will be needed in addition to information about its context. This is because different words can, of course, often occur in identical contexts, with the consequence that they would be classified as very similar on the basis of context alone.

Schütze (1995) compared the performance of an approach in which words were represented in terms of their context (as with, for example, Finch and Chater (1992a, 1992b)) with an approach in which words were represented in terms of their context *and* the context of neighbouring words. It was felt that the latter approach would provide important information about the syntactic properties of the words being considered.

The second type of representational vectors described above were clustered, after reduction to 50 dimensions using a singular value decomposition procedure, into 200 classes. The words in the corpus (the Brown corpus) were then replaced with a tag corresponding to the class into which they had been placed during the clustering procedure. On examining the words identified by particular tags, it was found that the

procedure did not always give satisfactory results because punctuation marks were included as part of the context. It was noted that the context vectors of punctuation marks conveyed little useful information about syntax, and for a further analysis words which neighboured punctuation marks were excluded from consideration. This modification to the technique permitted encouraging results to be obtained.

6.3 Conclusions

In Chapters 4 and 5 we examined the performance of a technique for producing a categorization of natural language in which each of a number of target words was represented by a statistical context vector. We noted there that rich structures could be obtained which were often in accordance with our intuitions about the similarity between the meanings of words, and that a comparison between the 'goodness' of the categorizations obtained under various conditions, using Roget's Thesaurus as a 'gold standard'. However, we have also noted that this type of 'standard' technique is handicapped by forcing only a single representation for each target word. Lexical ambiguity is a widespread feature of the English language, and it would therefore be desirable to permit more than one representation for each of the words considered, should the contexts in which the words occur dictate that this would be appropriate. Whilst this issue has not received a great deal of attention amongst researchers using corpus-based techniques to produce syntactic or semantic categorizations of language, we have seen that a number of investigations have nonetheless been carried out within the domain of Computational Linguistics which seek to allow words to be represented by more than one vector, and which present techniques for disambiguating between senses of a particular word token on the basis of its context.

Whilst these approaches are very much of relevance to the problem to be confronted here, they are generally not methods which are developed *on-line*. That is, they preserve another disadvantage of the 'standard' methods considered earlier in that they require separate stages of first constructing vectors and then measuring the distances between these. One of the objectives of the present thesis is to examine not only the extent to which the statistical structure of natural language can be used to categorize words on the basis of meaning, but the extent to which such a

categorization can be *developed*. Ideally, we would like to be able to present the system with a large sample of natural language from the outset and allow it to develop its own representations for each word-sense without any external supervision, and without the need for any separate stages of vector construction or statistical analysis. One possible solution to the problem is to incorporate the whole procedure within an unsupervised neural network, and it is to this that we shall now direct our attention.

7. *NEURAL NETWORK ANALYSES*

7.1 Elman's Approach to Syntactic Clustering

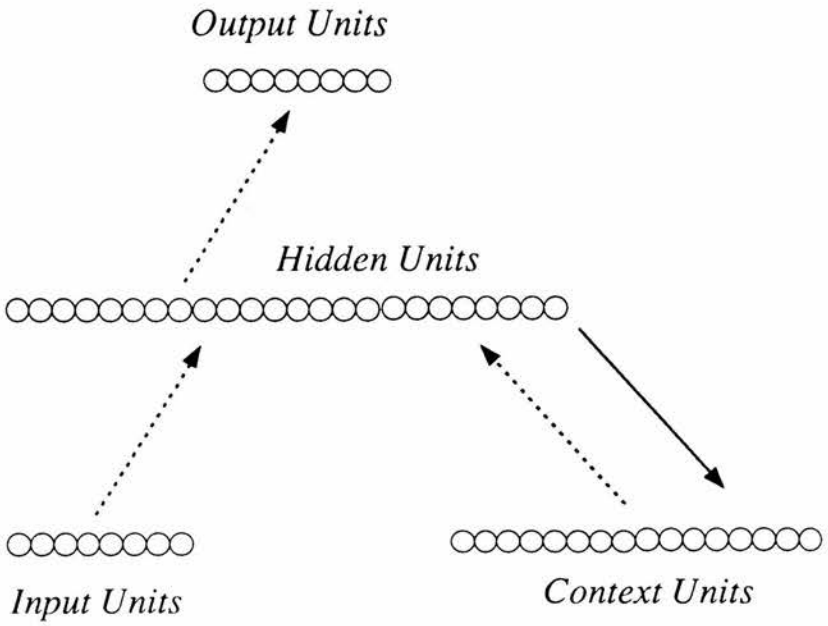
Elman (1988) provides an artificial corpus which, whilst containing only extremely simple grammatical structure, is ideal for the purpose of testing the capability of the neural network approach described in depth later in this chapter. Since we shall be considering performance using Elman's corpus in some detail, and since Elman himself used a neural network for the purposes of analysing this corpus, we shall now outline his approach.

Elman's motivation was one of enabling systems which use parallel processing, such as neural networks, to represent time. This, of course, is an important consideration when dealing with language, in which structure often appears as a temporal sequence. Elman's important intuition here was to allow a neural network to represent time *implicitly*, rather than explicitly as a dimension of the input to the network.

To realize this objective, Elman used a supervised neural network based on one described by Jordan in 1986. This is a recurrent network trained using the backpropagation algorithm (Rumelhart, Hinton and Williams, 1986). Time is able to exert its effect on the processing of the network by means of 'context units', which store information about *previous* states of the hidden units. The network architecture is outlined in figure 7.1 below.

At each time step, a segment of the language data is presented to the network. The input units and the hidden units both activate the output units, as is conventionally the case for a supervised multi-layer network. However, the hidden units are activated both by the input units and by the context units, which contain the hidden unit values from the previous time step. The recurrent connections, which feed from the hidden units to the context units, each have a fixed weight of 1 and are not subject to learning.

Figure 7.1: Outline of Elman's (1988) Supervised Recurrent Network



Since every input to the network is accompanied by information about the state of the hidden units at the previous time step, the network is able to learn about the temporal context of the inputs with which it is presented.

Elman applied a network of this type to various problems involving sequential input data. Of particular interest in the present context is Elman's consideration of the network's performance in discovering lexical classes from simple linguistic input. Elman generated a corpus of 27,354 words containing 10,000 sentence frames each of two or three words in length. The words were taken from a set of 29 lexical items[16], and each of these was represented by a randomly chosen 10-bit binary vector. The vectors were then concatenated to produce an input stream of 27,354 vectors. Sentence boundaries were not represented in any way.

The network used for this task was of the general type outlined above, having 10 input units, 10 output units, and a hidden layer of 50 units. There were also 50 context units.

[16] Whilst Elman (1988) makes mention of the use of 35 unique lexical items, details of only 29 such items appear in the paper.

The problem for the network was to learn to predict the next word in the input sequence from the current input. Training was continued for 5 passes through the 27,354 word sequence, giving a total of 136,770 training iterations. At the end of this training period, the average sum-squared error over the output units indicated that the network was not performing particularly well in the prediction task. However, training was terminated at this point to examine whether the network had learned anything about the categories to which the words belonged, using the information provided by the word order of the input sequence.

The network's performance in this regard was examined by looking at the internal representations it had developed. The weights in the network were frozen, and the entire 27,354 word sequence was presented again. This time, the hidden unit activation for each occurrence of each word was recorded, resulting in 27,354 50-bit vectors. The 50-bit vectors corresponding to each lexical item were then averaged to give just one 50-bit vector for each word. When these representations for each word were subjected to a hierarchical cluster analysis procedure, the resulting dendrogram revealed that the network had discovered several of the major categories of words in the corpus. Large categories were found which corresponded to 'nouns' and 'verbs', and within these the hierarchical structure revealed appropriate subcategories. Within the noun group, for example, subgroups were found for 'animates' (itself containing 'aggressors', 'small animals', and 'humans') and 'inanimates', and within the verb group subgroups were found for 'transitives' and 'intransitives'. In addition to these categories, groups for 'edibles' and 'breakables' were also found.

On the basis of these findings, Elman was able to conclude that :

"The network has developed internal representations for the input vectors which reflect facts about the possible sequential order of the inputs. The network is not able to predict the precise order of words, but it recognizes that (in this corpus) there is a class of inputs (viz., verbs) which typically follow other inputs (viz., nouns). This knowledge of class behavior is quite detailed ... (Elman 1988, p18)"

Elman also showed that when a particular word ('man') was replaced throughout the corpus with a novel word ('zog'), the internal representation developed for this word bore the same relationship to other words in the corpus as did the word it replaced.

Thus, since the linguistic context for the two words was similar, the hidden representations developed to represent them were also similar.

7.2 The Role of Distributional Statistics in Elman's Approach

The artificial corpus used by Elman (1988) contains a small number of words which fall into syntactic and semantic groupings on the basis of their word ordering as determined by a simple grammar. By using a recurrent neural network trained using the backpropagation algorithm, Elman was able to show that the network could learn to develop high level representations for the words which reflected these groupings.

The corpus used by Elman is very suitable for examining the performance of the unsupervised neural network introduced in this chapter. As with real samples of natural language, the corpus contains semantic groupings which can be discovered, given some information about the relationships between the words in the corpus.

However, we must first be aware of some more recent examinations of Elman's findings which provide further insights into the work being carried out by his network.

Chater and Conkey (1992) considered Elman's approach, which they described as an instance of a 'copy-back' training procedure. They repeated his analyses, but used as inputs a completely localist 1-of-n coding scheme, which required 29 input units to represent the 29 lexical items in the corpus[17]. Their network also contained 150 hidden units and 150 context units. For the purposes of analysis, Elman's approach was initially followed, and average hidden unit activation patterns were obtained for each word, resulting in a single 150 element vector for each unique word.

[17] Chater and Conkey (1992) used a 1-of-n coding scheme to represent the inputs, and the same approach is used in the analyses presented using the unsupervised network discussed later in this chapter. The advantage of this means of coding is that all input representations are guaranteed to be equidistant from each other in the input space, requiring the network to learn the relationships between them using information not present in the input representation itself. Elman's (1988) approach, on the other hand, could allow some input patterns to be closer together to each other than to others, albeit on a randomly selected basis.

When this analysis was carried out on the basis of hidden unit representations for the current input word only (as with Elman), relatively poor clustering resulted. However, Chater and Conkey were able to provide improved clustering results by averaging hidden unit representations on the basis of the word *predicted*, and also by averaging the *change* in the hidden unit representation brought about by a word. Thus, Chater and Conkey were able to produce dendrograms containing linguistically interesting categories from a variety of measures of hidden unit values.

Chater and Conkey assert that the reason for this is that each of the measures they considered corresponds to statistics within the corpus. Elman's original approach would correspond to grouping words by the conditional probabilities of successive words. This was tested empirically by measuring these conditional probabilities directly and then subjecting them to cluster analysis. It was found that very similar results to those produced by Chater and Conkey's neural network could be obtained in this way. It was also found that the improved clusterings obtained by different hidden unit measures noted above (based on the word predicted, or on the change in the hidden unit representation brought about by a word) could be closely approximated by using appropriate statistical analogues of these measures (clustering on the basis of the conditional probabilities of the preceding words, or the change in conditional probabilities expected after a word is input).

Chater and Conkey concluded that the limitation on the performance of a neural network in Elman's situation is the statistical structure of the data itself, rather than the nature of the network employed:

"These results suggest that the hidden unit patterns that recurrent neural networks develop can be viewed as reflecting quite directly the statistical structure of the sequences learnt. Furthermore, particular statistical measures of hidden unit activation may closely correspond to a related statistic of the sequence itself (p407)."

Thus, whatever the limitations of the recurrent network may have been here, the important role of statistical structure is once again underlined.

As a final remark on the status of Elman's network, it is worth bearing in mind the observation of Redington, Chater, and Finch (1995) that, since Elman's (1988)

analysis relies upon cluster analysis of the hidden unit activations, much of the computational work involved is not performed by the network itself. Furthermore, the approach is less economical than simply collecting the simple distributional statistics of the corpus; this requires only a single pass through the corpus, in contrast to the numerous passes required by Elman's network.

7.3 More Recent Neural Network Approaches

Following the upsurge in the use of neural network methods within Psychology and Cognitive Science in the mid 1980's, several psychological models concerning issues surrounding the acquisition of syntax and semantics were proposed (see, for example, Waltz and Pollack (1985), Rumelhart and McClelland (1986), and Seidenberg and McClelland (1989)).

Although many of these approaches employed supervised methods for training the neural networks involved, some researchers did consider the use of unsupervised networks in addressing such issues. Ritter and Kohonen (1989), for example, regarded such networks as providing a potential means for exploring the manner in which human linguistic categories might be learned from the linguistic input itself. As they have pointed out,

"... the internal representations of categories may be derivable from the mutual relations and roles of the primary signal or data elements themselves ... (p242)".

A Kohonen network was used to demonstrate this, and the linguistic input used was an artificial corpus containing examples of 498 different three-word sentences.

Each word was presented as input in the form of a separately calculated statistical vector representing its average context in the corpus over 10,000 sentences. The context for each word was considered only for word positions immediately adjacent to it.

After learning, the network was tested by presenting the words alone without context, and it was found that it had placed words with a similar meaning in neighbouring regions of the resulting 2-dimensional output map:

"... the contexts have "channeled" the word items to memory positions whose arrangement reflects both grammatical and semantic relationships (Ritter and Kohonen, 1989; p 249)".

Within the areas of the semantic map obtained, there was evidence of heirarchical structure; within a class of nouns, for example, separate areas were observed for words denoting types of food, words denoting animals, and words denoting proper names.

Scholtes (1991) also used a Kohonen network within the context of investigating language acquisition. He noted that the approach taken by Ritter and Kohonen (1989), and by Finch and Chater (1992a, 1992b) discussed below, does not provide a complete model for language acquisition because of the inability to learn language structure totally automatically; the network is not able to deal with sequential information. Instead, it requires information about each word's context to be summarized in the form of a vector. To deal with this shortcoming, Scholtes devised a network containing several layers of units connected both by feed-forward and feed-back connections. These connections were intended to inform particular units about the current input to the network and about the past input, thus providing information about context to these units. It was found that, after presenting various sentences to the network, that the architecture was successful in allowing the network to derive context for itself and to develop semantic maps like those discussed by Ritter and Kohonen (1989).

Finch and Chater (1992a, 1992b) followed their work using statistical methods for syntactic classification (discussed earlier in Chapter 3) with a method involving the use of an unsupervised neural network closely related to that of Ritter and Kohonen (1989). This could learn to represent words in terms of their distributional context, thus making use of the statistical structure discussed by Chater and Conkey (1992). Once the network had learned this kind of representation for the words, it was expected that similar words ought to be assigned similar representations, and that the network ought then to be able to find syntactic categories among these by cluster analysing the representations using the Kohonen algorithm for unsupervised clustering.

The input to Finch and Chater's network used a localist 1-of-n coding scheme, in which each of the 2000 input units would correspond to one of the 2000 unique words in the input data. A middle layer of units was provided in four banks, corresponding to the word at four neighbouring positions in the text (previous word but one, previous word, next word, next word but one). Each of these banks contained only 150 units, since only 150 context words were considered. The network was presented with a 40,000,000 word corpus taken from the Usenet. The weights between the input layer and the middle layer of the network were trained using simple Hebbian learning with normalization. Once training was completed, presentation of each input word would cause the representation in the middle layer of the network to reflect the distribution of contexts in which the input word had occurred. These middle layer representations were then clustered into 100 groups using a Kohonen network. These groups revealed that the network was able to cluster words into groups sharing the same syntactic category, although it was also found that more than one of the resulting clusters would correspond to the same syntactic category.

Schütze (1993a) has also used a neural network approach for classifying words in large corpora on the basis of syntax. As with the unsupervised approach to be introduced below, Schütze (1993a) was concerned with the problem of lexical ambiguity in performing such analyses. A recurrent supervised network similar to that used by Elman (1988) was employed. However, vectors resulting from a singular value decomposition of co-occurrence vectors were used as inputs and targets. In addition, the network was described as *birecurrent*, having recurrency both to the left and to the *right* of the target word. The input to the network at each time step consisted of the word to the left of the target, the left context of the target at the previous time step, the word to the right of the target, and the right context of the target at the next time step. During testing, the network was required to predict the syntactic category of the target word presented at the input units. The results obtained here were found to be promising, even where ambiguous words were concerned (that is, those word tokens belonging to more than one syntactic category).

Schütze (1993b) also considered the problem, directly relevant to this thesis, of classifying words on the basis of semantics using neural networks. Instead of representing words in terms of their co-occurrence statistics with other words, the co-occurrences between 'letter fourgrams' were used. The reason for taking this step was to reduce the number of zero or very low, unreliable co-occurrence counts which would inevitably arise when recording co-occurrences between a large number of words from a corpus. Since the fourgrams chosen were frequent ones, it was expected that the resulting co-occurrence counts ought to be relatively reliable. For each target word considered, a vector of co-occurrences with 5000 fourgrams was constructed. As with many such analyses, the context vector representing each target words was summed over all occurrences of the target word, resulting in a single representation for the target word which did not distinguish between word senses. The only exception to this was that a distinction was made between target words beginning with an upper case letter and those beginning with a lower case letter. The representations used were felt to be useful as potential input representations for neural networks; removal of various components of the vectors resulted in a graceful degradation in performance, indicating that the representations were of a distributed nature.

Random samples of the target words revealed that their nearest neighbours in the context vector space were often words semantically related to them. Not surprisingly, however, it was found that the neighbours of words which were used in a wide variety of contexts in the corpus were not so close to them in meaning.

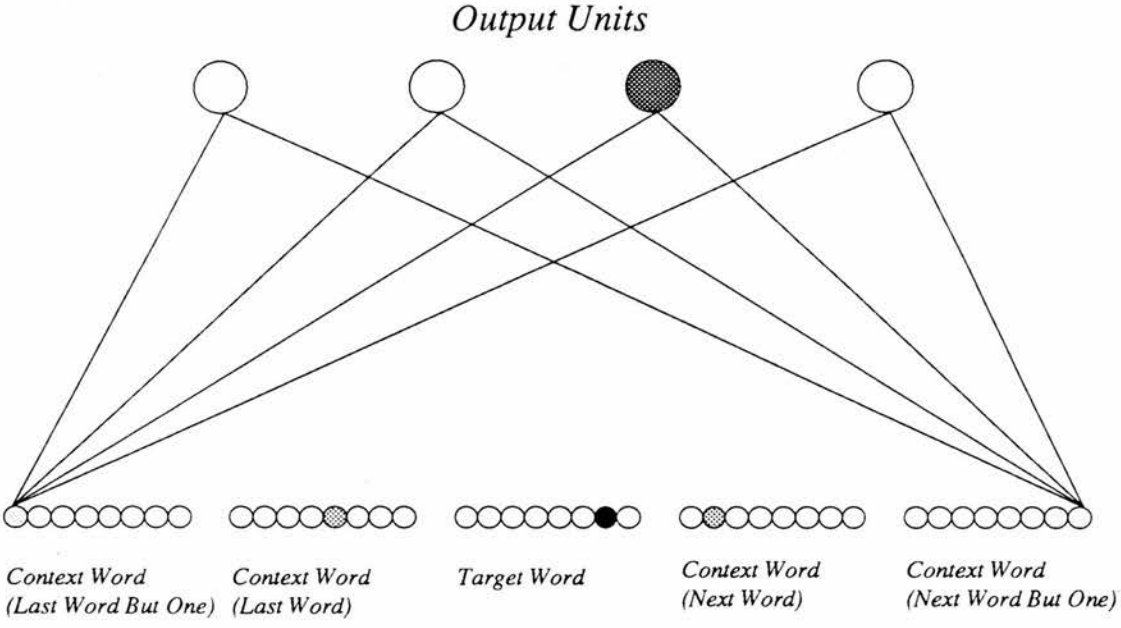
7.4 A New Approach

A new neural network approach, initially outlined in Huckle (1995), will now be discussed. This seeks to satisfy our original requirement for an unsupervised system to perform a classification of language, but which also attempts to overcome some of the limitations of the 'standard' approaches discussed in Chapter 6, and which aims to improve upon the neural network approaches which have hitherto been applied to the problem of language categorization.

The performance of the network will first be considered on the corpus devised by Elman (1988), involving a comparison with the standard approaches on this corpus, and then its performance will be considered on samples of the Wall Street Journal corpus.

The overall architecture of the network is outlined in figure 7.2 below.

Figure 7.2: Outline of Unsupervised Neural Network for Clustering Target Words



The input layer of the network uses a localist coding scheme, as did Finch and Chater's (1992a, 1992b) network. However, instead of presenting only the target word as input, and presenting the context words to a middle layer of the network, both target words and context words are presented at the input layer in this network, with the intention of allowing it to learn about context sequentially, rather than having this provided to the network in the form of a separately calculated vector, as with Ritter and Kohonen (1989) and Finch and Chater (1992a, 1992b). The network proposed here, then, approaches the problem in a similar manner to Scholtes (1991), albeit with a much less complex architecture. This is rather less unorthodox than Finch and Chater's approach in which some of the input data (namely the context words) are presented to the middle layer of a network.

The approach here also allows context words to be presented along with the target word during the test phase, as well as during the training phase. If Finch and Chater had wished to present context information during the test phase, their network would have required context to be presented to the middle layer during this phase, just as it was during the training phase. In fact, this was not of interest to them because at the test stage they wished to perform clustering over the middle layer representations which arose from presentation of the target word alone, these having been developed by Hebbian learning during the training phase of their network.

Being able to present context words during the test phase seems important because we will very often be interested in what the network makes of a word given its present context, rather than what it makes of a word presented in isolation, as with Finch and Chater (1992a, 1992b) and Ritter and Kohonen (1989). For these researchers, clustering was carried out on the basis of a 'smearing' over all the word's contexts during the training phase. The ability to represent context is likely to be crucial in attempting to disambiguate words presented during the test phase of the network.

With the network described here, there is no major distinction between the training phase and the test phase of the network other than the cessation of learning in the latter. In Finch and Chater's case, the two phases are quite different; Hebbian learning stops, input data is no longer presented to the middle layer of the network, and Kohonen clustering commences along the lines of the approach used by Ritter and Kohonen (1989).

The input units of the network are, in a similar fashion to Finch and Chater's network, composed of a number of banks of units. Each bank uses 1-of-n coding to represent the words concerned. In the centre of the input pattern is a bank of units representing the target word, and to either side are banks of units representing the context words to the left and right of the target word. The network is capable of representing an arbitrary number of target words and an arbitrary number of context words at an arbitrary number of positions from the target word. Furthermore, unequal numbers of such positions can be accommodated (we could, for example, examine performance

using context words at 4 positions to the left of the target word and 1 position to the right of the target word). The network is fully connected, with each input unit connecting to each output unit.

In order to allow the network to distinguish between the target word and the context words (which is important since the target word is what we wish to categorize, and to do this successfully, as Schütze (1995) has pointed out, requires information about the word itself as well as its context), the activation pattern over the input units is greatest for the bank of units representing the target word, and is less for the banks of units representing the context words, decreasing with their distance from the target word. This characteristic of the input representation is similar to the 'dynamic context vector' proposed by Gallant (1991), which gave decreasing weight to context as a function of its distance from the target word. Indeed, Gallant noted that his proposals for the use of a context vector approach in performing word sense disambiguation would be appropriate for a neural network implementation very much like the one proposed here. The decreasing activation in this case was arranged using by setting the activation for the active unit in each bank of input units at

$$\xi_j = 1/(D+1) \ , \qquad (7.1)$$

where D is the distance between the target word and the input in question. Thus, the context input at a position immediately adjacent to the target word will give D a value of 1, and thus the activation of this input is set to 1/2. Similarly, the context input at the next-but-one position will give D a value of 2, and thus will have an activation of 1/3, and so on.

This also seems reasonable from the point of view that human language learners would be expected to pay more attention to the word currently being heard than to words surrounding it. As Marslen-Wilson (1989), for example, points out, word recognition when listening to speech is an on-line process in which words are, on average, recognized in context about 200 milliseconds after onset. This suggests that,

whilst context is of great importance in word recognition, the word currently being heard plays an especially significant role in this process.

The output units perform a type of 'winner-take-all' clustering of the input patterns (see, for example, Rumelhart and Zipser (1985) for an extended discussion of this technique). This clustering starts at the beginning of the training phase of the network and continues until learning ceases. The rationale behind this is that, in the human case, it might not be reasonable to suppose that categorization begins once some kind of representation has been built up to capture the contexts in which a word occurs (as with many of the techniques we have discussed so far in this thesis), but that it might be more reasonable to suppose that such categorization starts from the beginning. Of course, it will at first be likely to give rise to imperfect categories, but it might be expected that these will evolve over time to produce more satisfactory ones. In any case, we have the possibility with the network described of being able to terminate training at any point and examining the clusters developed so far.

The weights in the network are updated using a version of the standard competitive learning rule (Hertz, Krogh, and Palmer, 1991). Of the output units λ_i, the 'winning' unit λ_{i^*} is defined as the one having the largest net input

$$h_i = \sum_j w_{ij} \xi_j^{\mu} \qquad (7.2)$$

where ξ_j^{μ} is the activation of the *j*th element of the current input vector ξ^{μ}, and w_{ij} is the weight from the *j*th element of the input vector to the *i*th output unit.

The weights w_{ij} are initially set to randomly chosen values between 0 and 1, and normalized throughout using the $\sum_j w_{ij} = 1$ normalization for each output unit λ_i.

The normalization has the consequence that the winning output unit will be the one with the weight vector closest in direction to that of the input vector. Thus, for each input presentation, the output unit with the 'sense' most closely corresponding to that

of the input vector should be the winning unit. This has clear parallels with the approach advocated by Gallant (1991) and explored by Schütze (1992).

The weight vector $\mathbf{w}_{i^*j}$ for the winning output unit λ_{i^*} is updated using the rule

$$\Delta w_{i^*j} = \eta(\xi_j^\mu - w_{i^*j}) \qquad (7.3),$$

and the weight vectors $\mathbf{w}_{ij}$ for the other output units are updated using the rule

$$\Delta w_{ij} = \varphi(\xi_j^\mu - w_{ij}) \qquad (7.4),$$

where $\varphi < \eta$.

Thus, the weight vector for the winning output unit is shifted towards the current input vector by an amount more than for the other output units. It is not strictly necessary that the non-winning output units undergo any weight change, but it is a means by which 'dead units' which never win for any inputs can be avoided and can gradually be brought into use. The learning rules above are also more successful if the input vectors are normalized using the $\sum_j \xi_j^\mu = 1$ normalization.

It was foreseen that a problem might arise in training the network to cluster its inputs because some input patterns would be likely to occur much more often than others. Given Zipf's law, some target words and their contexts are likely to occur much more often than others, meaning that some input patterns may dominate the competitive learning process. In the 'standard' analyses discussed in Chapter 4, this kind of problem is removed by normalizing for the frequency of the target words. In the present case, the intention was to develop a system which could work 'on-line' and with as few *post-hoc* adjustments as possible. Therefore, the problem of frequency was tackled by arranging that the network would pay gradually less and less attention to target words as they occurred more and more often. Thus, common words would rapidly have little effect on learning, whilst the less frequent words would continue to

bring about a greater amount of learning on the occasions when they occurred. To achieve this, the learning rules 7.2 and 7.3 were modified slightly, so that in practice the weight vector $\mathbf{w}_{i^*j}$ for the winning output unit λ_{i^*} was updated using the rule

$$\Delta w_{i^*j} = \alpha\eta(\xi_j^{\mu} - w_{i^*j}) \qquad (7.5),$$

and the weight vectors $\mathbf{w}_{ij}$ for the other output units were updated using the rule

$$\Delta w_{ij} = \alpha\varphi(\xi_j^{\mu} - w_{ij}) \qquad (7.6).$$

In these rules, the parameter α is equal to the reciprocal of the frequency of the target word being presented to the network at the time. For large amounts of input text, this means that there should be only very small differences in the amount of learning brought about by target words of high and low frequency. The policy of paying less attention to inputs which have been encountered frequently in the past, and more attention to inputs which are novel, is often an important one for neural networks[18]. Murre, Phaf, and Wolters (1992), for example, have made use of a similar means of dealing with the frequency of inputs in an unsupervised neural network used for categorization, and they note that such an approach has a firm biological precedent:

"The reduction in learning rate and random activations with repeated presentation of a pattern may be compared to habituation found in almost all organisms. Presentation of an unfamiliar stimulus to an animal gives rise to an arousal response, which may be accompanied by an orientation reaction ... But repetition of the same stimulus results in a gradual habituation of the response (p60)".

For the purposes of evaluation and comparison of this network, let us first examine the performance of the standard approach, used in Chapter 4, on the Elman data. This will be carried out for a number of sets of parameters to give a reasonable illustration of the performance of the 'standard' analysis in dealing with Elman's corpus.

[18] The issue here has been described by Grossberg (1987), for example, as the Stability-Plasticity dilemma; the system should remain stable when irrelevant events are encountered, yet should be plastic enough to be able to adapt to significant events when they are encountered.

7.5 The Elman Grammar

A corpus of 27354 lexical items conforming to Elman's artificial grammar was generated using software kindly supplied by Martin Redington[19]. Elman's corpus places the 29 lexical items into the following syntactic categories:

Table 7.1: Syntactic Categories Contained in Elman's (1988) Corpus

Category	Members
NOUN-HUM	man, woman, girl, boy
NOUN-ANIM	cat, dog, mouse
NOUN-INANIM	book, rock, car
NOUN-AGRESS	dragon, lion, monster
NOUN-FOOD	cookie, bread, sandwich
NOUN-FRAG	plate, glass
VERB-EAT	eat
VERB-AGPAT	move
VERB-DESTROY	break, smash
VERB-TRAN	like, chase
VERB-PERCEPT	smell, see
VERB-INTRAN	think, sleep, exist

7.6 Analyses

7.6.1 Analysis 1

The parameters used for this analysis are listed below in table 7.2.

Table 7.2: Parameters Used in Analysis 1

Corpus	Elman (1988)
Number of Words in Corpus	27354
Window Length[20]	1
Number of Target Words Considered	29
Number of Context Words Used	29
Distance Metric	Spearman Rank Correlation Coefficient

For each of the 29 items in the corpus, the nearest neighbours were calculated using the Spearman coefficient, and these are presented in table 7.3.

[19] Martin Redington is at the Department of Experimental Psychology, University of Oxford.

[20] Here, and elsewhere in this chapter, the window length refers to the number of words enclosed within the moving window *on each side* of the target word.

Table 7.3: Nearest Neighbours for the Target Words Considered in Analysis 1

Target Word	Nearest Neighbours (Spearman Correlation Coefficient)
book	car (0.979) rock (0.973) exist (0.634) sleep (0.624) think (0.624) chase (0.567) like (0.563) move (0.498) cookie (0.472) bread (0.467) see (0.464) sandwich (0.461) smell (0.458) smash (0.451) break (0.450) eat (0.370) plate (0.327) glass (0.321) mouse (0.121) cat (0.099) dog (0.091) girl (-0.152) woman (-0.163) monster (-0.225) lion (-0.248) dragon (-0.289) boy (-0.294) man (-0.298)
boy	girl (0.665) man (0.589) woman (0.489) dragon (0.345) dog (0.333) monster (0.312) lion (0.302) mouse (0.258) cat (0.232) see (0.121) smell (0.116) chase (-0.029) like (-0.033) smash (-0.213) break (-0.224) rock (-0.253) move (-0.282) car (-0.284) book (-0.294) sandwich (-0.314) cookie (-0.317) bread (-0.329) eat (-0.401) sleep (-0.453) glass (-0.459) exist (-0.478) think (-0.478) plate (-0.482)
bread	sandwich (0.987) cookie (0.983) sleep (0.883) exist (0.881) think (0.878) move (0.750) chase (0.738) like (0.730) plate (0.660) glass (0.652) smell (0.632) see (0.631) break (0.541) smash (0.540) eat (0.496) rock (0.471) cat (0.471) book (0.467) car (0.461) mouse (0.458) dog (0.415) monster (0.042) lion (0.012) dragon (-0.025) girl (-0.047) woman (-0.101) man (-0.299) boy (-0.329)
break	smash (0.996) think (0.620) exist (0.613) sleep (0.596) bread (0.541) sandwich (0.524) cookie (0.503) chase (0.480) eat (0.476) see (0.466) smell (0.466) like (0.465) plate (0.462) book (0.450) glass (0.424) rock (0.379) car (0.378) move (0.357) dog (0.175) cat (0.151) mouse (0.149) monster (-0.042) lion (-0.065) dragon (-0.101) boy (-0.224) woman (-0.268) man (-0.273) girl (-0.300)
car	book (0.979) rock (0.972) exist (0.626) sleep (0.625) think (0.607) like (0.605) chase (0.603) move (0.519) cookie (0.477) sandwich (0.465) bread (0.461) see (0.450) smell (0.441) eat (0.381) break (0.378) smash (0.376) glass (0.324) plate (0.316) cat (0.132) mouse (0.117) dog (0.089) woman (-0.130) girl (-0.165) monster (-0.188) lion (-0.220) dragon (-0.235) boy (-0.284) man (-0.309)
cat	mouse (0.914) dog (0.874) chase (0.549) like (0.548) cookie (0.494) bread (0.471) sandwich (0.465) girl (0.412) smell (0.411) see (0.408) sleep (0.356) think (0.351) move (0.342) exist (0.340) woman (0.335) boy (0.232) lion (0.221) eat (0.211) monster (0.196) dragon (0.173) smash (0.162) rock (0.157) break (0.151) car (0.132) man (0.128) book (0.099) glass (0.081) plate (0.052)
chase	like (0.997) sleep (0.818) exist (0.799) think (0.789) cookie (0.772) sandwich (0.744) bread (0.738) glass (0.719) move (0.718) plate (0.682) rock (0.611) see (0.609) smell (0.608) car (0.603) book (0.567) eat (0.558) cat (0.549) smash (0.491) mouse (0.488) break (0.480) dog (0.474) woman (0.217) monster (0.193) lion (0.188) dragon (0.157) girl (0.138) man (0.060) boy (-0.029)
cookie	bread (0.983) sandwich (0.979) sleep (0.882) exist (0.880) think (0.876) chase (0.772) like (0.771) move (0.767) glass (0.670) plate (0.662) see (0.633) smell (0.629) smash (0.508) break (0.503) cat (0.494) eat (0.481) car (0.477) rock (0.476) mouse (0.475) book (0.472) dog (0.449) monster (0.080) lion (0.032) dragon (0.006) woman (-0.036) girl (-0.051) man (-0.243) boy (-0.317)
dog	mouse (0.928) cat (0.874) chase (0.474) like (0.468) cookie (0.449) smell (0.445) sandwich (0.445) see (0.433) girl (0.427) bread (0.415) move (0.366) woman (0.344) sleep (0.336) boy (0.333) exist (0.323) think (0.318) monster (0.254) lion (0.227) smash (0.200) man (0.191) break (0.175) dragon (0.175) eat (0.133) rock (0.120) book (0.091) car (0.089) glass (0.046) plate (0.022)
dragon	lion (0.960) monster (0.955) boy (0.345) girl (0.323) see (0.305) smell (0.298) man (0.249) dog (0.175) cat (0.173) mouse (0.168) like (0.163) chase (0.157) woman (0.113) glass (0.100) move (0.079) plate (0.066) cookie (0.006) sandwich (-0.008) bread (-0.025) eat (-0.062) smash (-0.093) break (-0.101) sleep (-0.149) exist (-0.163) think (-0.167) car (-0.235) rock (-0.242) book (-0.289)
eat	exist (0.596) think (0.591) sleep (0.588) like (0.560) chase (0.558) sandwich (0.500) bread (0.496) cookie (0.481) break (0.476) smash (0.467) plate (0.407) move (0.406) glass (0.393) car (0.381) rock (0.376) book (0.370) smell (0.236) see (0.232) cat (0.211) mouse (0.173) dog (0.133) lion (0.020) dragon (-0.062) monster (-0.085) girl (-0.283) woman (-0.317) boy (-0.401) man (-0.476)
exist	sleep (0.994) think (0.985) bread (0.881) cookie (0.880) sandwich (0.879) move (0.860) chase (0.799) like (0.793) plate (0.781) glass (0.775) see (0.698) smell (0.696) book (0.634) car (0.626) rock (0.622) smash (0.614) break (0.613) eat (0.596) mouse (0.353) cat (0.340) dog (0.323) monster (-0.083) lion (-0.129) dragon (-0.163) girl (-0.193) woman (-0.212) man (-0.387) boy (-0.478)
girl	boy (0.665) man (0.660) woman (0.617) mouse (0.456) dog (0.427) cat (0.412) lion (0.392) smell (0.371) see (0.351) monster (0.330) dragon (0.323) chase (0.138) like (0.123) bread (-0.047) move (-0.050) sandwich (-0.050) cookie (-0.051) glass (-0.072) plate (-0.082) rock (-0.107) book (-0.152) car (-0.165) sleep (-0.174) think (-0.176) exist (-0.193) smash (-0.279) eat (-0.283) break (-0.300)
glass	plate (0.991) exist (0.775) sleep (0.772) think (0.767) chase (0.719) like (0.717) move (0.672) cookie (0.670) sandwich (0.653) bread (0.652) see (0.564) smell (0.560) smash (0.428) break (0.424) eat (0.393) car (0.324) rock (0.321) book (0.321) monster (0.171) lion (0.126) dragon (0.100) cat (0.081) mouse (0.069) dog (0.046) girl (-0.072) woman (-0.123) man (-0.166) boy (-0.459)
like	chase (0.997) sleep (0.810) exist (0.793) think (0.782) cookie (0.771) sandwich (0.736) bread (0.730) move (0.721) glass (0.717) plate (0.678) rock (0.608) car (0.605) see (0.597) smell (0.593) book (0.563) eat (0.560) cat (0.548) mouse (0.488) smash (0.475) dog (0.468) break (0.465) woman (0.232) monster (0.201) lion (0.187) dragon (0.163) girl (0.123) man (0.058) boy (-0.033)
lion	monster (0.961) dragon (0.960) girl (0.392) smell (0.317) see (0.310) boy (0.302) mouse (0.237) dog (0.227) cat (0.221) man (0.213) chase (0.188) like (0.187) woman (0.162) glass (0.126) plate (0.104) move (0.062) cookie (0.032) eat (0.020) sandwich (0.012) bread (0.012) smash (-0.053) break (-0.065) sleep (-0.115) think (-0.126) exist (-0.129) rock (-0.196) car (-0.220) book (-0.248)

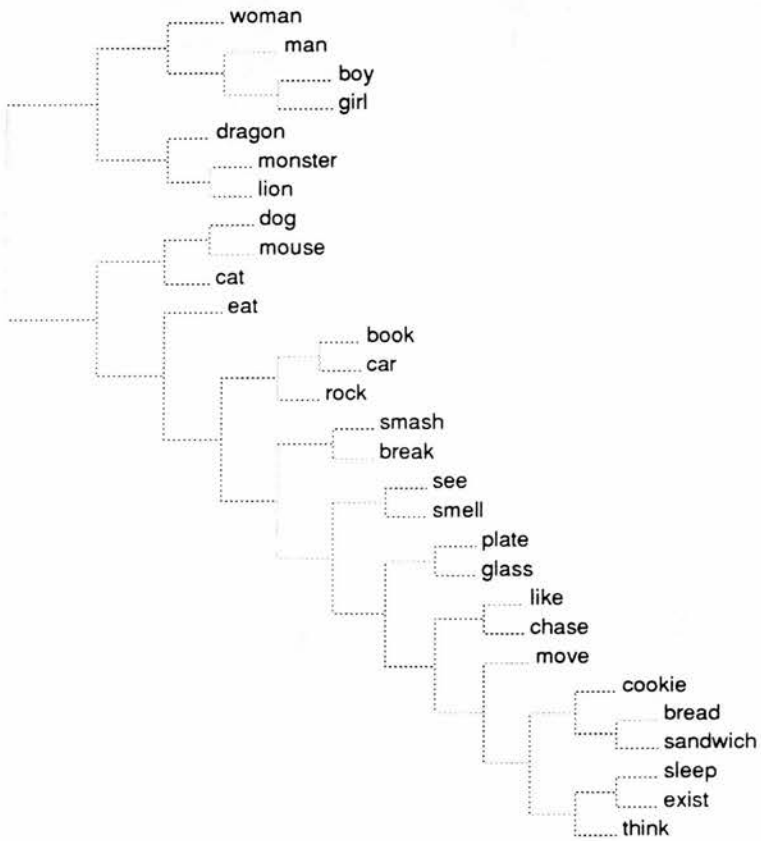
Table 7.3 (contd.)

man	girl (0.660) woman (0.597) boy (0.589) dragon (0.249) monster (0.219) lion (0.213) dog (0.191) mouse (0.172) cat (0.128) see (0.088) smell (0.073) chase (0.060) like (0.058) glass (-0.166) plate (-0.205) cookie (-0.243) smash (-0.256) sandwich (-0.261) break (-0.273) move (-0.295) book (-0.298) bread (-0.299) car (-0.309) rock (-0.320) exist (-0.387) sleep (-0.399) think (-0.409) eat (-0.476)
monster	lion (0.961) dragon (0.955) see (0.366) smell (0.363) girl (0.330) boy (0.312) dog (0.254) mouse (0.240) man (0.219) like (0.201) cat (0.196) chase (0.193) woman (0.173) glass (0.171) move (0.149) plate (0.145) cookie (0.080) sandwich (0.046) bread (0.042) smash (-0.035) break (-0.042) sleep (-0.075) exist (-0.083) eat (-0.085) think (-0.097) rock (-0.184) car (-0.188) book (-0.225)
mouse	dog (0.928) cat (0.914) chase (0.488) like (0.488) smell (0.483) see (0.482) cookie (0.475) bread (0.458) girl (0.456) sandwich (0.445) move (0.363) exist (0.353) sleep (0.350) think (0.342) woman (0.332) boy (0.258) monster (0.240) lion (0.237) eat (0.173) man (0.172) dragon (0.168) smash (0.160) break (0.149) rock (0.145) book (0.121) car (0.117) glass (0.069) plate (0.065)
move	sleep (0.863) exist (0.860) think (0.857) see (0.780) smell (0.775) cookie (0.767) sandwich (0.758) bread (0.750) like (0.721) chase (0.718) glass (0.672) plate (0.655) rock (0.521) car (0.519) book (0.498) eat (0.406) dog (0.366) mouse (0.363) smash (0.360) break (0.357) cat (0.342) monster (0.149) dragon (0.079) lion (0.062) girl (-0.050) woman (-0.191) boy (-0.282) man (-0.295)
plate	glass (0.991) exist (0.781) think (0.775) sleep (0.771) chase (0.682) like (0.678) cookie (0.662) bread (0.660) move (0.655) sandwich (0.649) see (0.567) smell (0.566) smash (0.463) break (0.462) eat (0.407) book (0.327) rock (0.319) car (0.316) monster (0.145) lion (0.104) dragon (0.066) mouse (0.065) cat (0.052) dog (0.022) girl (-0.082) woman (-0.149) man (-0.205) boy (-0.482)
rock	book (0.973) car (0.972) think (0.630) sleep (0.626) exist (0.622) chase (0.611) like (0.608) move (0.521) cookie (0.476) bread (0.471) sandwich (0.467) smell (0.465) see (0.462) smash (0.389) break (0.379) eat (0.376) glass (0.321) plate (0.319) cat (0.157) mouse (0.145) dog (0.120) woman (-0.097) girl (-0.107) monster (-0.184) lion (-0.196) dragon (-0.242) boy (-0.253) man (-0.320)
sandwich	bread (0.987) cookie (0.979) sleep (0.886) exist (0.879) think (0.870) move (0.758) chase (0.744) like (0.736) glass (0.653) plate (0.649) smell (0.618) see (0.617) smash (0.529) break (0.524) eat (0.500) rock (0.467) car (0.465) cat (0.465) book (0.461) dog (0.445) mouse (0.445) monster (0.046) lion (0.012) dragon (-0.008) girl (-0.050) woman (-0.096) man (-0.261) boy (-0.314)
see	smell (0.995) move (0.780) exist (0.698) sleep (0.696) think (0.693) cookie (0.633) bread (0.631) sandwich (0.617) chase (0.609) like (0.597) plate (0.567) glass (0.564) mouse (0.482) smash (0.476) break (0.466) book (0.464) rock (0.462) car (0.450) dog (0.433) cat (0.408) monster (0.366) girl (0.351) lion (0.310) dragon (0.305) eat (0.232) woman (0.138) boy (0.121) man (0.088)
sleep	exist (0.994) think (0.983) sandwich (0.886) bread (0.883) cookie (0.882) move (0.863) chase (0.818) like (0.810) glass (0.772) plate (0.771) smell (0.697) see (0.696) rock (0.626) car (0.625) book (0.624) smash (0.600) break (0.596) eat (0.588) cat (0.356) mouse (0.350) dog (0.336) monster (-0.075) lion (-0.115) dragon (-0.149) girl (-0.174) woman (-0.193) man (-0.399) boy (-0.453)
smash	break (0.996) think (0.630) exist (0.614) sleep (0.600) bread (0.540) sandwich (0.529) cookie (0.508) chase (0.491) smell (0.480) see (0.476) like (0.475) eat (0.467) plate (0.463) book (0.451) glass (0.428) rock (0.389) car (0.376) move (0.360) dog (0.200) cat (0.162) mouse (0.160) monster (-0.035) lion (-0.053) dragon (-0.093) boy (-0.213) woman (-0.247) man (-0.256) girl (-0.279)
smell	see (0.995) move (0.775) think (0.698) sleep (0.697) exist (0.696) bread (0.632) cookie (0.629) sandwich (0.618) chase (0.608) like (0.593) plate (0.566) glass (0.560) mouse (0.483) smash (0.480) break (0.466) rock (0.465) book (0.458) dog (0.445) car (0.441) cat (0.411) girl (0.371) monster (0.363) lion (0.317) dragon (0.298) eat (0.236) woman (0.162) boy (0.116) man (0.073)
think	exist (0.985) sleep (0.983) bread (0.878) cookie (0.876) sandwich (0.870) move (0.857) chase (0.789) like (0.782) plate (0.775) glass (0.767) smell (0.698) see (0.693) smash (0.630) rock (0.630) book (0.624) break (0.620) car (0.607) eat (0.591) cat (0.351) mouse (0.342) dog (0.318) monster (-0.097) lion (-0.126) dragon (-0.167) girl (-0.176) woman (-0.214) man (-0.409) boy (-0.478)
woman	girl (0.617) man (0.597) boy (0.489) dog (0.344) cat (0.335) mouse (0.332) like (0.232) chase (0.217) monster (0.173) smell (0.162) lion (0.162) see (0.138) dragon (0.113) cookie (-0.036) sandwich (-0.096) rock (-0.097) bread (-0.101) glass (-0.123) car (-0.130) plate (-0.149) book (-0.163) move (-0.191) sleep (-0.193) exist (-0.212) think (-0.214) smash (-0.247) break (-0.268) eat (-0.317)

It should be clear from inspection of this table that the nearest neighbours for the items in the corpus are usually those which share the same category in the original grammar. The only exceptions occur with 'eat' and 'move', which each belong to their own category; in these cases, the nearest neighbours are nonetheless ones from an intuitively related category.

These findings are in accordance with the findings of Chater and Conkey (1992). Since the number of lexical items involved is small, and the results presented here are so straightforward, we shall not attempt any further quantitative assessment of this analysis other than to present a dendrogram constructed in the same manner as those discussed in Chapter 4 (see figure 7.3).

Figure 7.3: Dendrogram Resulting From Analysis 1



Again, it is clear that the clustering provides a categorization in close agreement with the original grammatical categories used by Elman (1988), and one which is very similar to that resulting from the cluster analysis of the average hidden unit activations in the neural network he used for the analysis.

7.6.2 Analysis 2

The parameters used for this analysis are listed below in table 7.4.

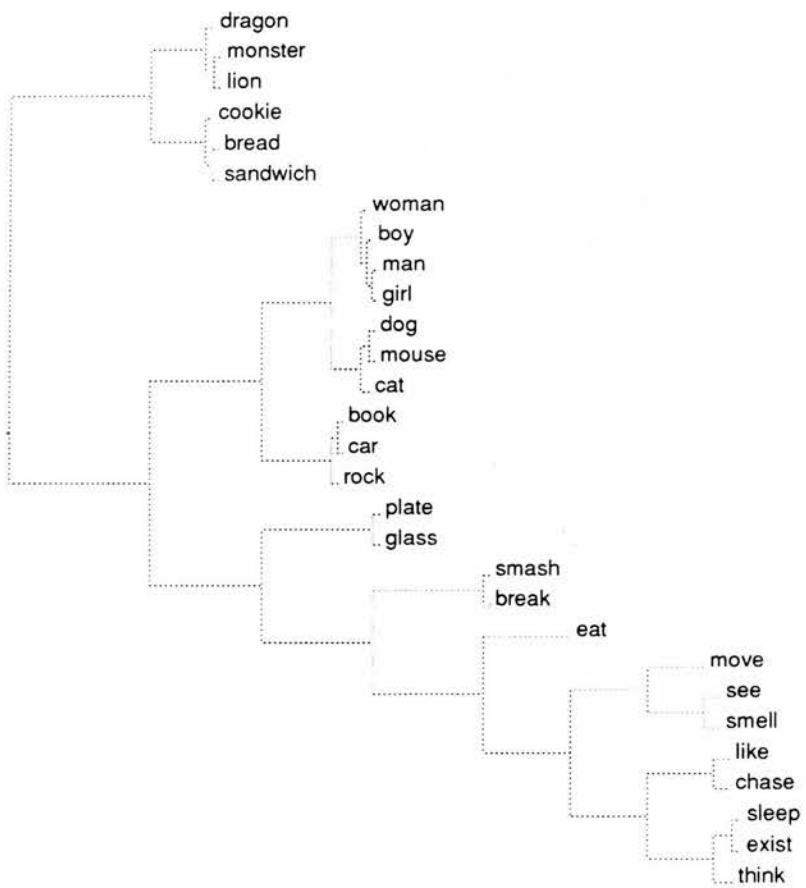
Table 7.4: Parameters Used in Analysis 2

Corpus	Elman (1988)
Number of Words in Corpus	27354
Window Length	1
Number of Target Words Considered	29
Number of Context Words Used	29
Distance Metric	Euclidean Distance

The table of nearest neighbours for analysis 2 may be found in Appendix B, table B.1.

The dendrogram resulting from this analysis is shown below in figure 7.4.

Figure 7.4: Dendrogram Resulting From Analysis 2



7.6.3 Analysis 3

The parameters used for this analysis are listed below in table 7.5.

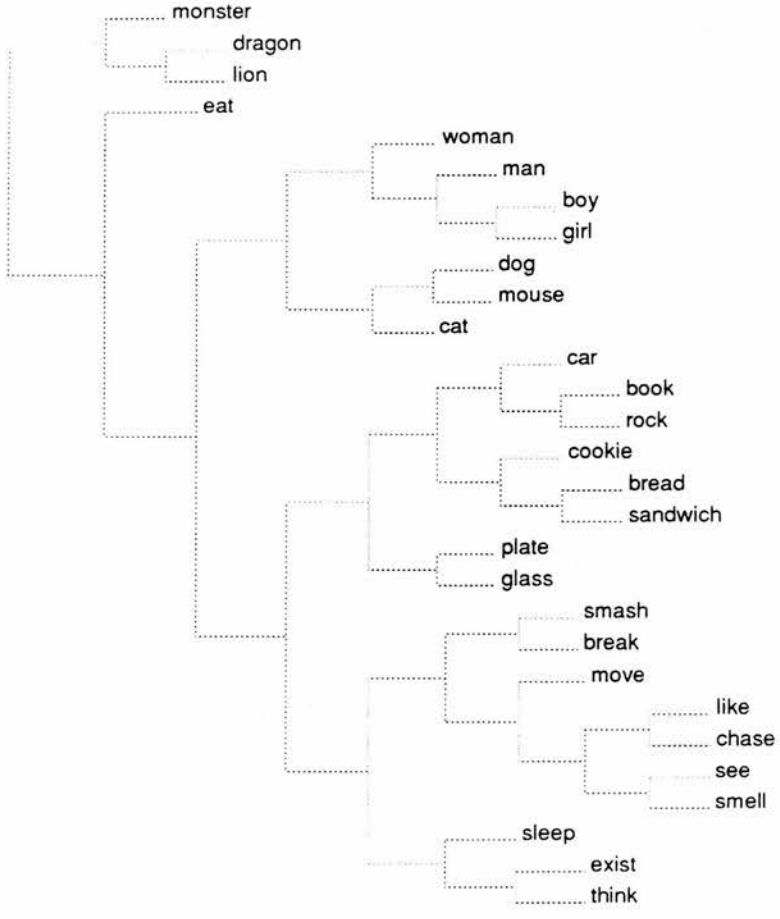
Table 7.5: Parameters Used in Analysis 3

Corpus	Elman (1988)
Number of Words in Corpus	27354
Window Length	2
Number of Target Words Considered	29
Number of Context Words Used	29
Distance Metric	Spearman Rank Correlation Coefficient

The table of nearest neighbours for analysis 3 may be found in Appendix B, table B.2.

The dendrogram resulting from this analysis is shown below in figure 7.5.

Figure 7.5: Dendrogram Resulting From Analysis 3



7.6.4 Analysis 4

The parameters used for this analysis are listed below in table 7.6.

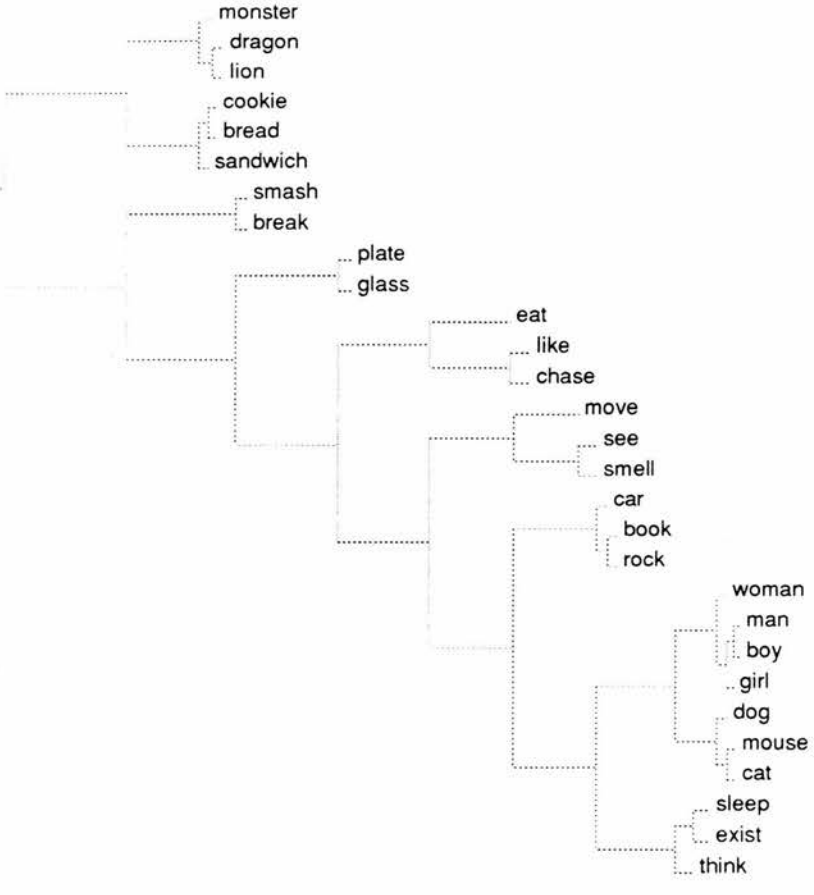
Table 7.6: Parameters Used in Analysis 4

Corpus	Elman (1988)
Number of Words in Corpus	27354
Window Length	2
Number of Target Words Considered	29
Number of Context Words Used	29
Distance Metric	Euclidean Distance

The table of nearest neighbours for analysis 4 may be found in Appendix B, table B.3.

The dendrogram resulting from this analysis is shown below in figure 7.6.

Figure 7.6: Dendrogram Resulting from Analysis 4



It is clear from the results of these analyses that the 'standard' analyses, in which target words are represented as vectors of conditional probabilities, do allow the statistical structure in Elman's (1988) corpus to be used for the purposes of categorizing the words it contains. This, of course, also confirms Chater and Conkey's (1992) findings.

Having confirmed that Elman's corpus provides a source of input data which is straightforward for the 'standard' analyses to deal with successfully, we can now turn our attention to the performance of the neural network described earlier.

7.7 Neural Network Analysis of Elman Data

Let us first examine the performance of the network when 2 output units are available to cluster the input data. This approach is, incidentally, of interest from a linguistic point of view because recent formulations of Chomsky's (1965) approach to syntax have emphasized the importance of binary branching structures in the acquisition of syntactic structures. Haegeman (1984), for example, motivates this on the grounds that the use of theories of phrase structure which permit only binary branching structures greatly reduces the number of permissible sentence structures. This constrained approach in turn means that a child learning language would have to make many fewer decisions when assigning syntactic structure to language data.

The parameters for the network were set as indicated in table 7.7 below:

Table 7.7: Parameters Used for Neural Network Analyses of Elman's (1988) Corpus

Corpus	Elman (1988)
Number of Target Words/Training Iterations	27354
Window Length	2
Number of Target Words Considered	29
Number of Context Words Used	29
Learning rate for winning unit	10
Learning rate for other unit(s)	1

The software written to simulate the network was provided with graphical output in X-Windows to permit a constant picture of its behaviour to be provided. This is important, since we are interested in the ability of the network to cluster the input words into coherent groups from the beginning, rather than waiting until training has finished. Samples of this graphical output will shortly be presented for the analyses carried out on the Elman data.

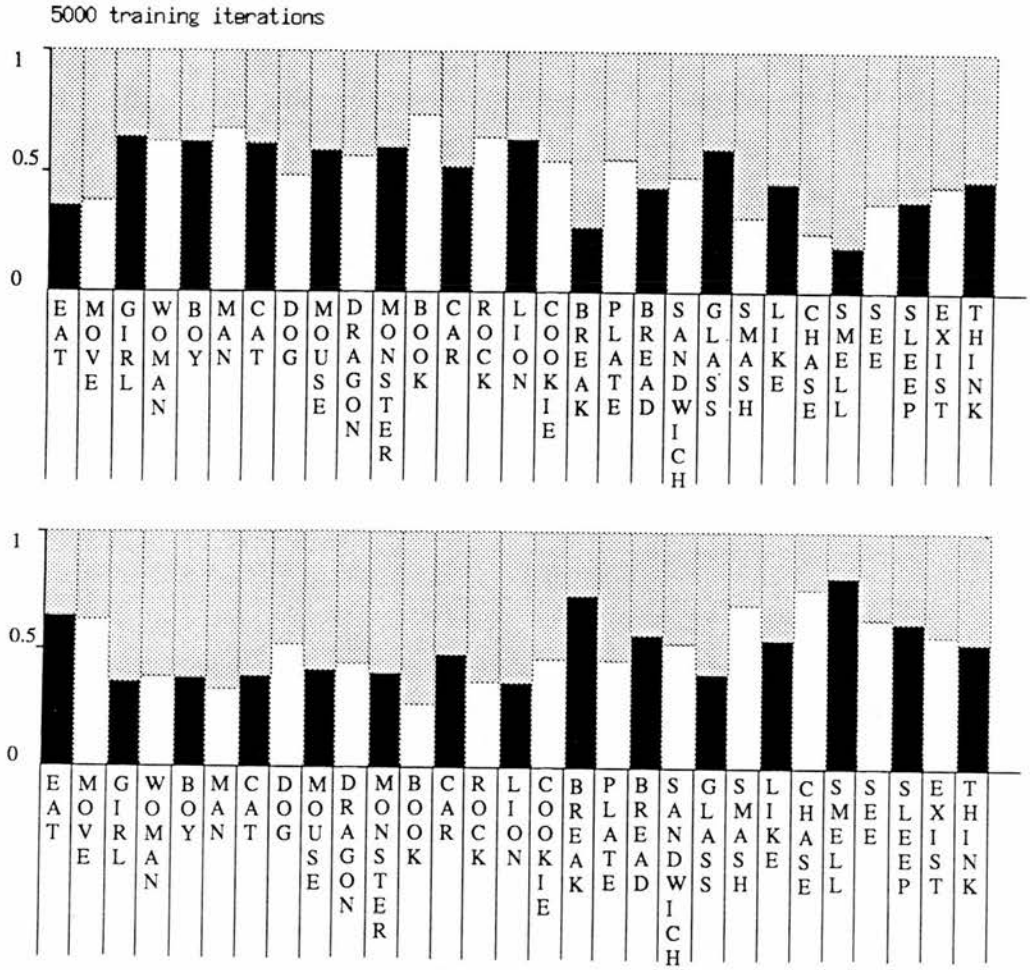
For each of the output units in the network, an indicator is constantly updated to show the probability with which that output unit will respond to each of the target words. This display can be updated at every training iteration to give a detailed impression of how the probabilities change as learning takes place, or can be updated at less frequent intervals as desired. For the Elman data, we shall see how the output units are behaving after every 5,000 training iterations until training is complete. Once the weights in the network have been frozen, and testing is under way, the display continues to show the appropriate probabilities. For each of the indicators used, the vertical axis indicates the probability of response, while the horizontal axis is used to represent the target words concerned.

The probability of each of the 2 units responding to each of the possible target words is now presented for different points during training. In each case, the graphs corresponding to the two output units are presented one above the other.

It is important to emphasize at this point the difference between the parameters of window length and of the number of output units in the network. The former parameter refers to the distance (in terms of words) around each target word within which context words form part of the input to the network. For instance, a window length of 2 indicates that context words within a distance of 2 words either side of the target word were used as input. This is, of course, quite different from the parameter concerned with the number of output units in the network.

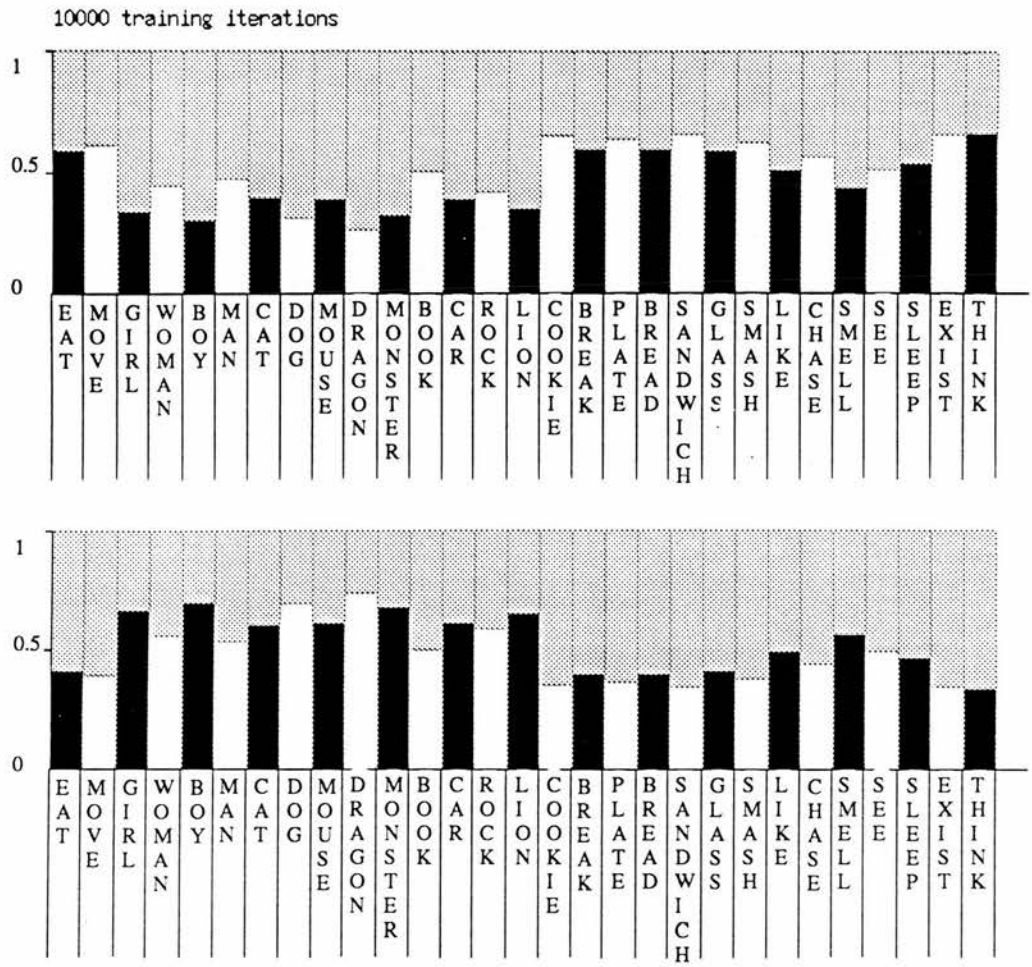
Figure 7.7 below provides the response data for the network after the first 5,000 training iterations. It can be seen that none of the probabilities is very close to 1 or 0, and that there is no obvious distinction between the words to which the two units will respond.

Figure 7.7: Output Unit Responses After 5,000 Training Iterations



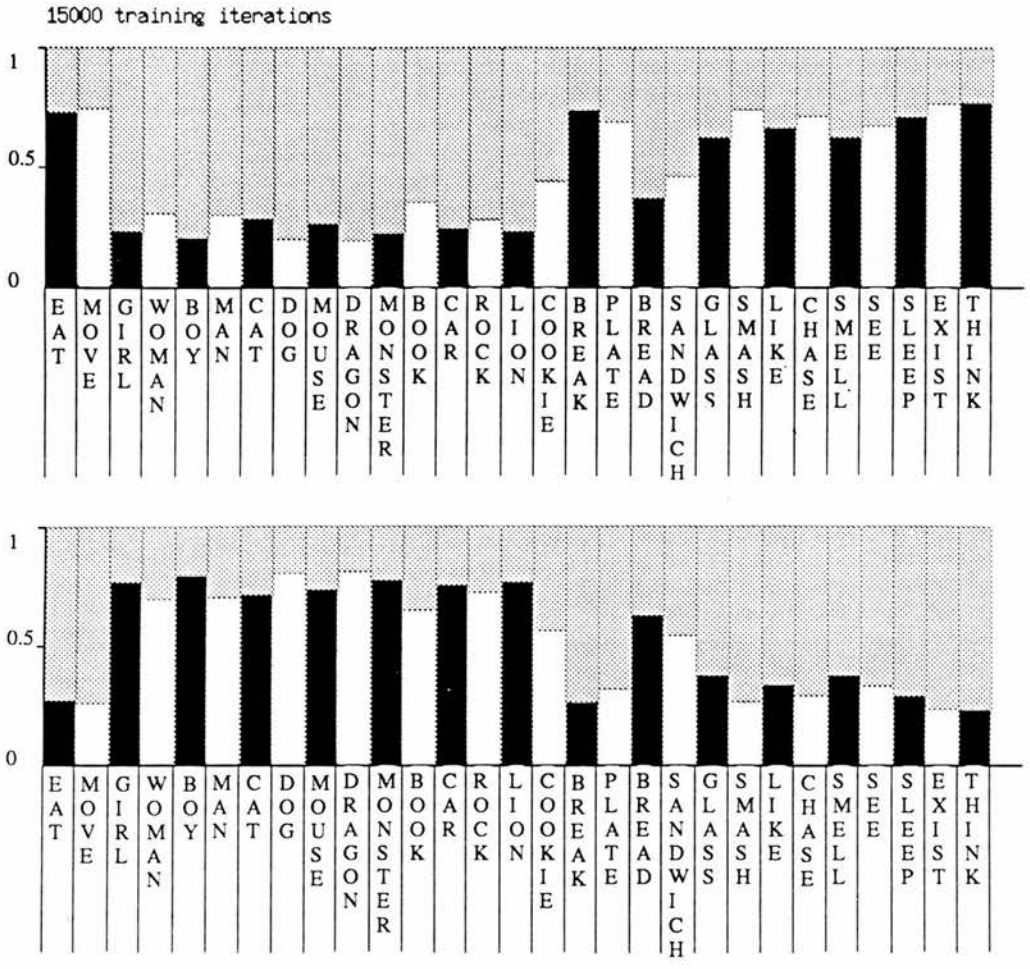
In figure 7.8 below, the picture is presented for the network after 10,000 training iterations. The situation is still unclear, but there is evidence of a tendency for the upper unit to be responding preferentially to the first two words, and to a larger group of words on the right of the graph. This situation is, of course, reversed for the lower unit since the two probabilities for each word must sum to one.

Figure 7.8: Output Responses After 10,000 Training Iterations



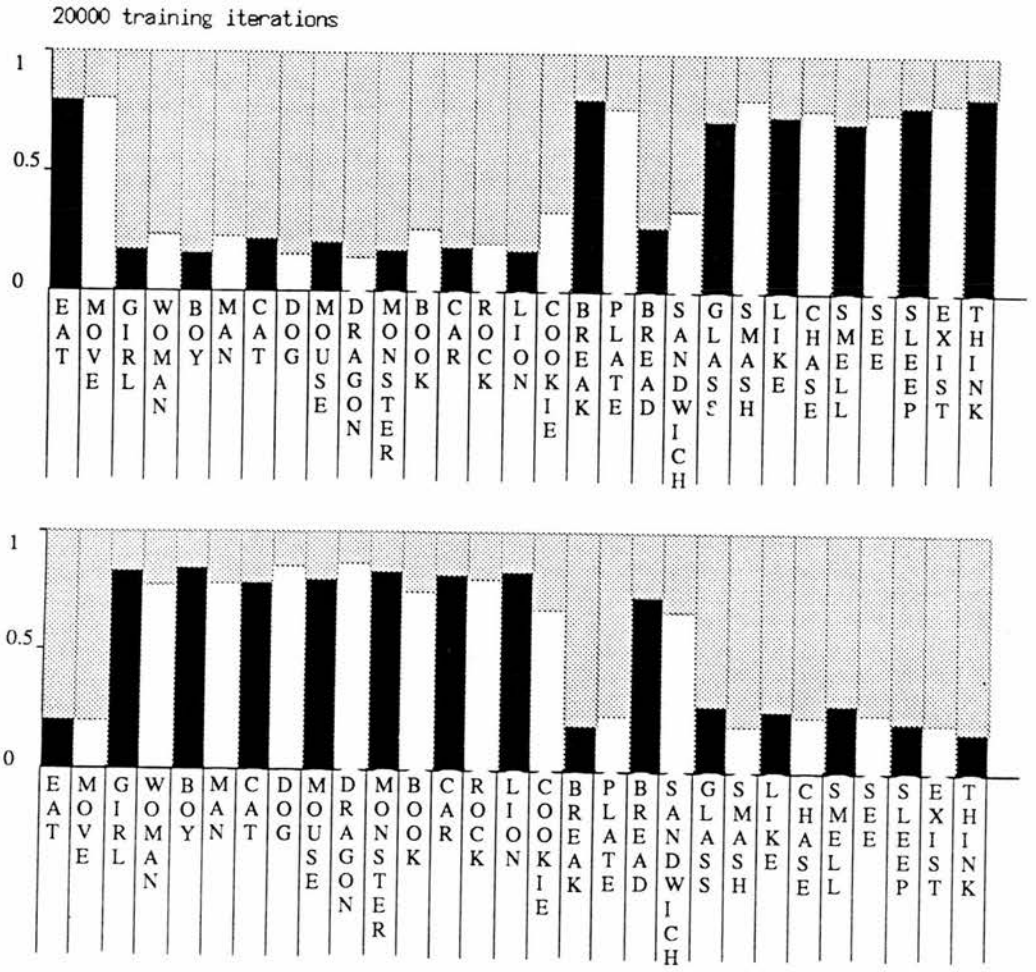
The behaviour of the network after 15,000 iterations, shown in figure 7.9 below, continues the trend seen at 10,000 training iterations. It is now clearly the case that each unit is responding much more often for some words than for others.

Figure 7.9: Output Unit Responses After 15,000 Training Iterations



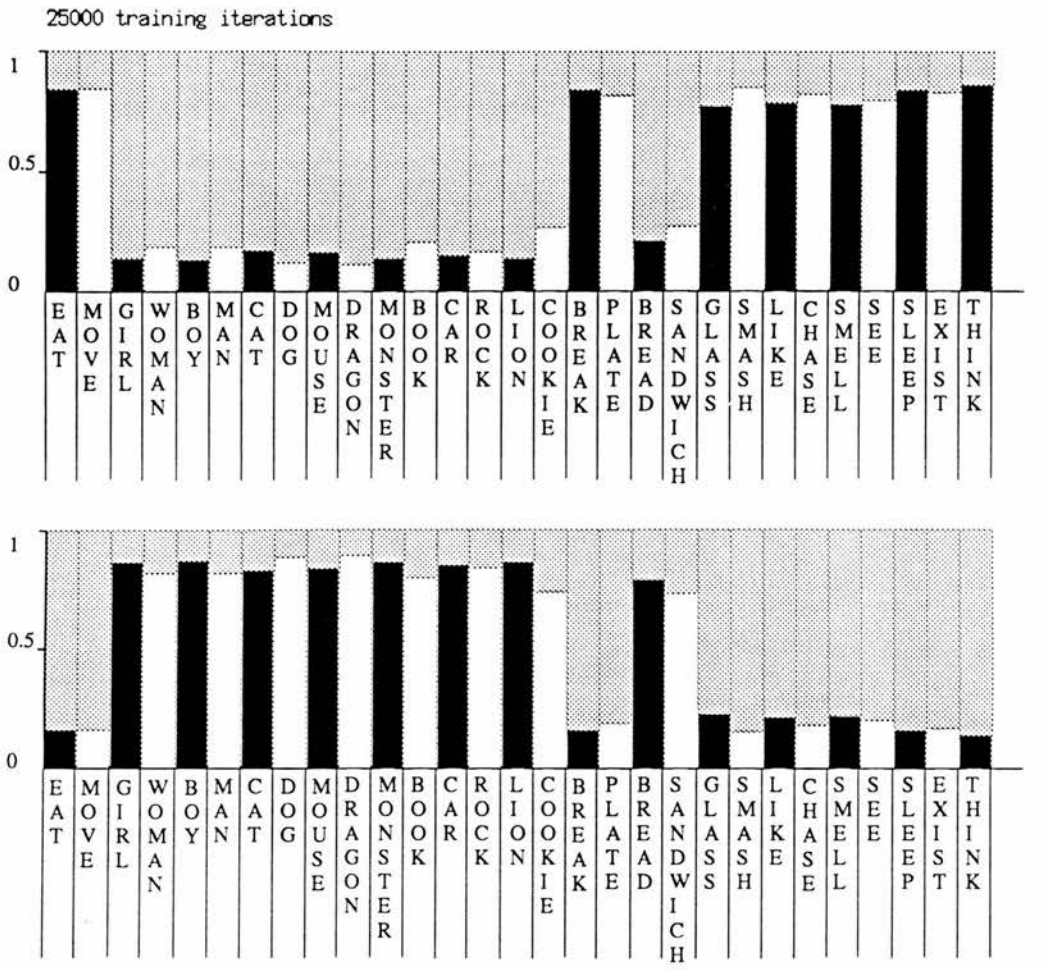
After 20,000 training iterations, the position is shown in figure 7.10. Again, the split between the two units' response patterns is becoming still more apparent.

Figure 7.10: Output Unit Responses After 20,000 Training Iterations



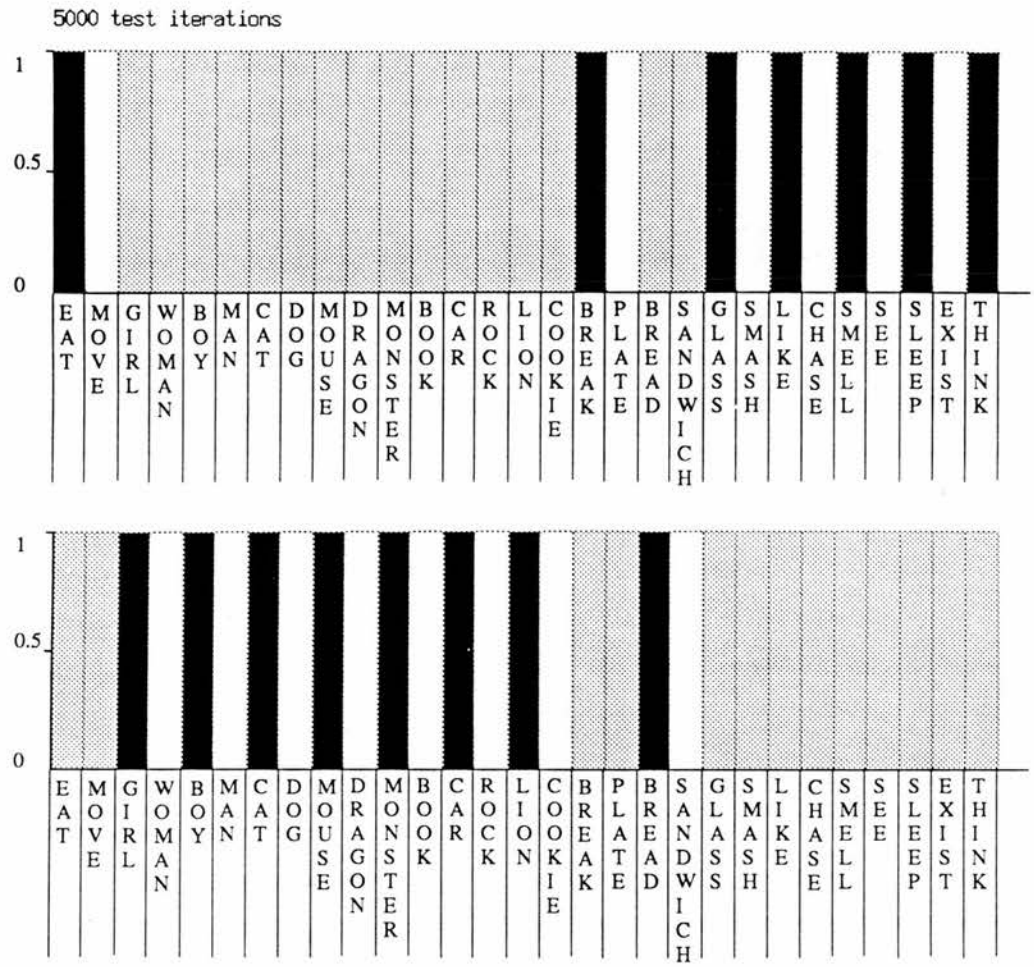
In figure 7.11 below, the situation is shown for the network after 25,000 training iterations, shortly before training ceases at 27,354 iterations.

Figure 7.11: Output Responses After 25,000 Training Iterations



After training was completed, the responses of the two units were recorded during the test phase, in which the network was presented with further text from the corpus used. After 5,000 test iterations, the responses of the two units used are as shown in figure 7.12 below. It can be seen that, following exposure to the corpus during training, the two units have learned to respond to the words in the corpus in an all-or-none fashion; all the probabilities have become 1 or 0.

Figure 7.12: Output Unit Responses After 5,000 Test Iterations



Furthermore, it is clear that the split between the two units is not a random one here; the first unit depicted is responding to all the verbs in the corpus, plus two of the nouns ('plate' and 'glass'), whilst the lower unit is responding to all the remaining nouns. So, using two output units to perform the clustering, the network has learned to distinguish almost perfectly between the two highest level categories in Elman's corpus; namely, nouns and verbs. The two nouns which are inappropriately clustered do appear to be on the borderline between the group of nouns and the group of verbs in the dendrograms presented earlier. This suggests that, statistically speaking, the context of these words is similar to that of the verbs and that the network's manner of dealing with them can thus be understood.

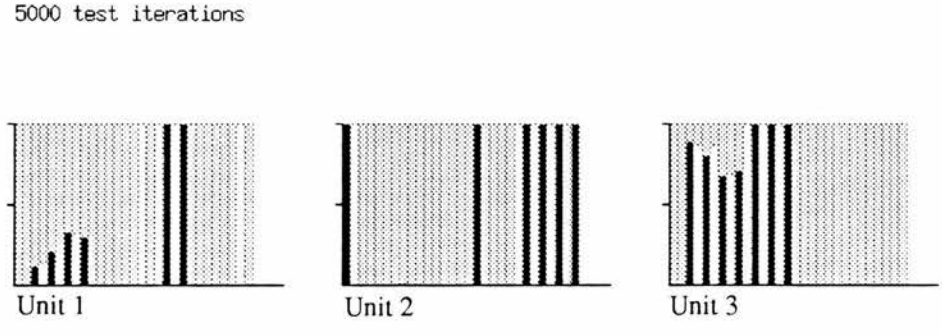
We may ask how the network would perform if it were allowed more than 2 output units. To answer this, the simulations were run again using a number of different output units to examine how the output units would distinguish between the target words in the corpus. To reduce the amount of space taken up by the presentation of these results, graphical displays are provided only for the position after 5,000 test iterations in each case. The displays are also presented in a smaller size, although the ordering of the words across the horizontal axis is the same as for the larger displays shown above. The parameters used in training the network are also the same as above apart from the number of output units used.

Table 7.8: Responses for Neural Network With 3 Output Units

Output Unit	Words Responded to Exclusively	Other Words Responded to
1	cookie, plate, bread, sandwich, glass, exist	girl, woman, boy, man, cat, dog, mouse
2	eat, move, break, see, sleep, think, smell, chase, like, smash	
3	dragon, monster, book, car, rock, lion	girl, woman, boy, man, cat, dog, mouse

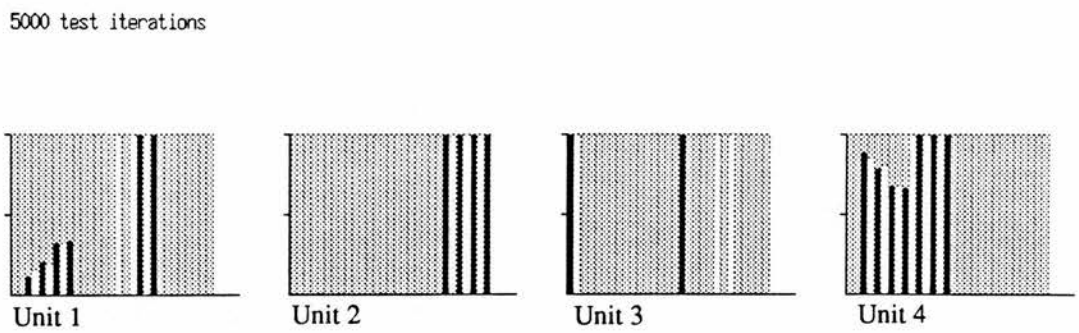
In figure 7.13 below, the response characteristics for each of 3 output units is shown. It can be seen that not all words are responded to exclusively by one of the output units; this was, of course, a capability desired from the outset to allow the network to deal with polysemy. The words to which each unit responds are summarized in table 7.8.

Figure 7.13: Output Unit Responses After 5,000 Test Iterations



Examination of figure 7.13 reveals that units 1 and 3 are responding almost entirely to nouns, whilst unit 2 is responding with a probability of 1 to all verbs other than 'exist' In figure 7.14 below, the situation is presented for the network with 4 output units.

Figure 7.14: Output Unit Responses After 5,000 Test Iterations



As in the previous analysis, it can be seen that some words are responded to by more than one output unit. The details are listed in table 7.9 below.

Table 7.9: Responses for Neural Network With 4 Output Units

Output Unit	Words Responded to Exclusively	Other Words Responded to
1	glass, sandwich, bread, plate, cookie	girl, woman, boy, man, cat, dog, mouse
2	think, exist, sleep, see, smell, like	rock
3	eat, move, chase, smash, break	
4	dragon, monster, book, car, lion	girl, woman, boy, man, cat, dog, mouse, rock

Again, clear distinctions are being made between the nouns and verbs in the corpus, with evidence of the original groupings of words in the grammar having been detected by the network.

Extending the number of output units further, figure 7.15 presents the response characteristics for the network with 5 output units.

Figure 7.15: Output Unit Responses After 5,000 Test Iterations

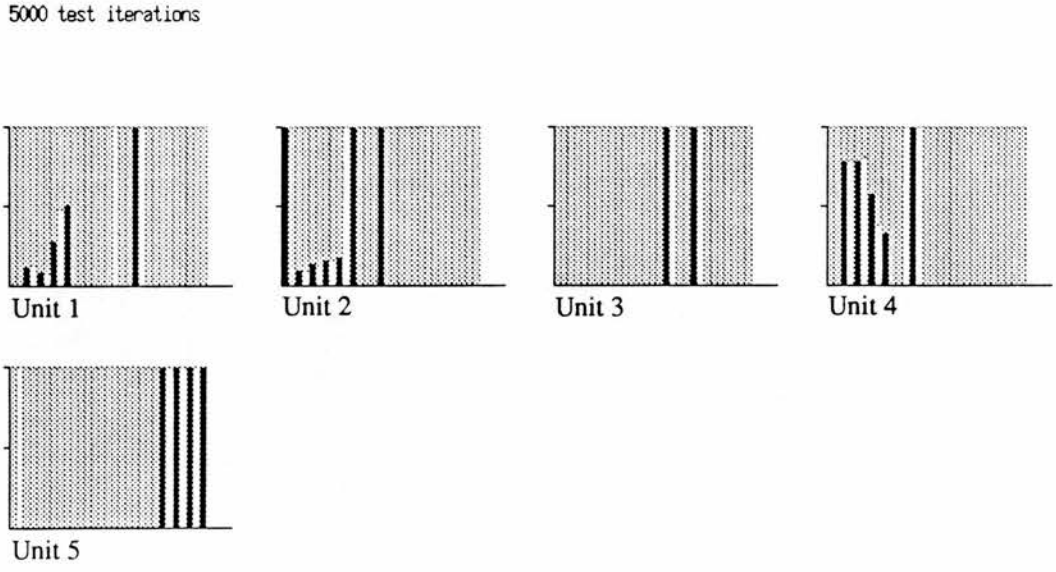


Table 7.10 indicates that, again, several of the lower level category distinctions present in the original corpus are being picked up by the different output units available in this analysis.

Table 7.10: Responses for Neural Network With 5 Output Units

Output Unit	Words Responded to Exclusively	Other Words Responded to
1	bread, sandwich, cookie	girl, woman, boy, man, cat, dog, mouse
2	eat, dragon, monster, lion	girl, woman, boy, man, cat, dog, mouse
3	glass, smash, plate, break	
4	book, car, rock	girl, woman, boy, man, cat, dog, mouse
5	move, like, chase, smell, see, sleep, exist, think	

The original corpus designed by Elman contained 12 syntactic categories (see table 7.1 above). Given this, it is of interest to see how well a network equipped with 12 output units can perform in recovering these categories. This was tested by running a simulation with 12 output units (in principle, one unit for each of the categories in the corpus), and the results are presented in figure 7.16 below.

Figure 7.16: Output Unit Responses After 5,000 Test Iterations

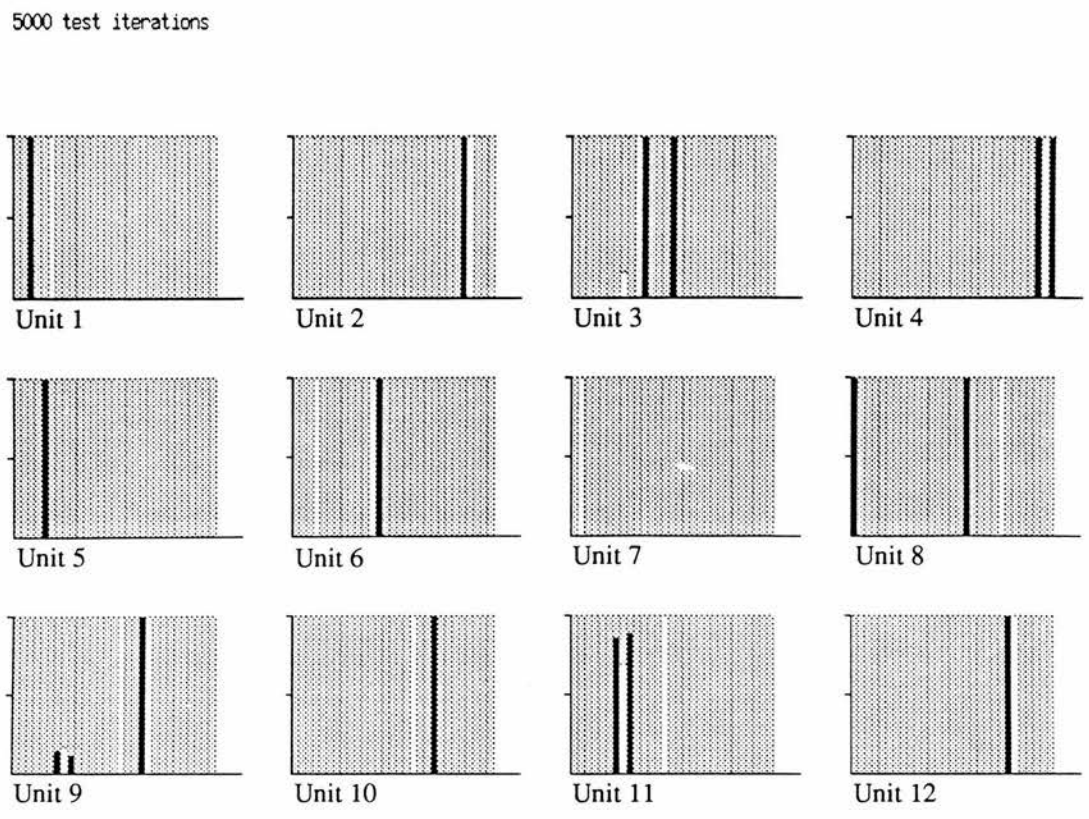


Table 7.11: Responses for Neural Network with 12 Output Units

Output Unit	Words Responded to Exclusively	Other Words Responded to	Elman Category
1	girl, man		
2	see, smell		VERB-PERCEPT
3	dragon, monster, lion	dog	
4	sleep, exist, think		VERB-INTRAN
5	boy		
6	woman, book, car		
7	move		VERB-AGPAT
8	eat, break, smash		
9	cookie, bread, sandwich	cat, dog, mouse	
10	glass, plate		NOUN-FRAG
11	rock	cat, dog, mouse	
12	like, chase		VERB-TRAN

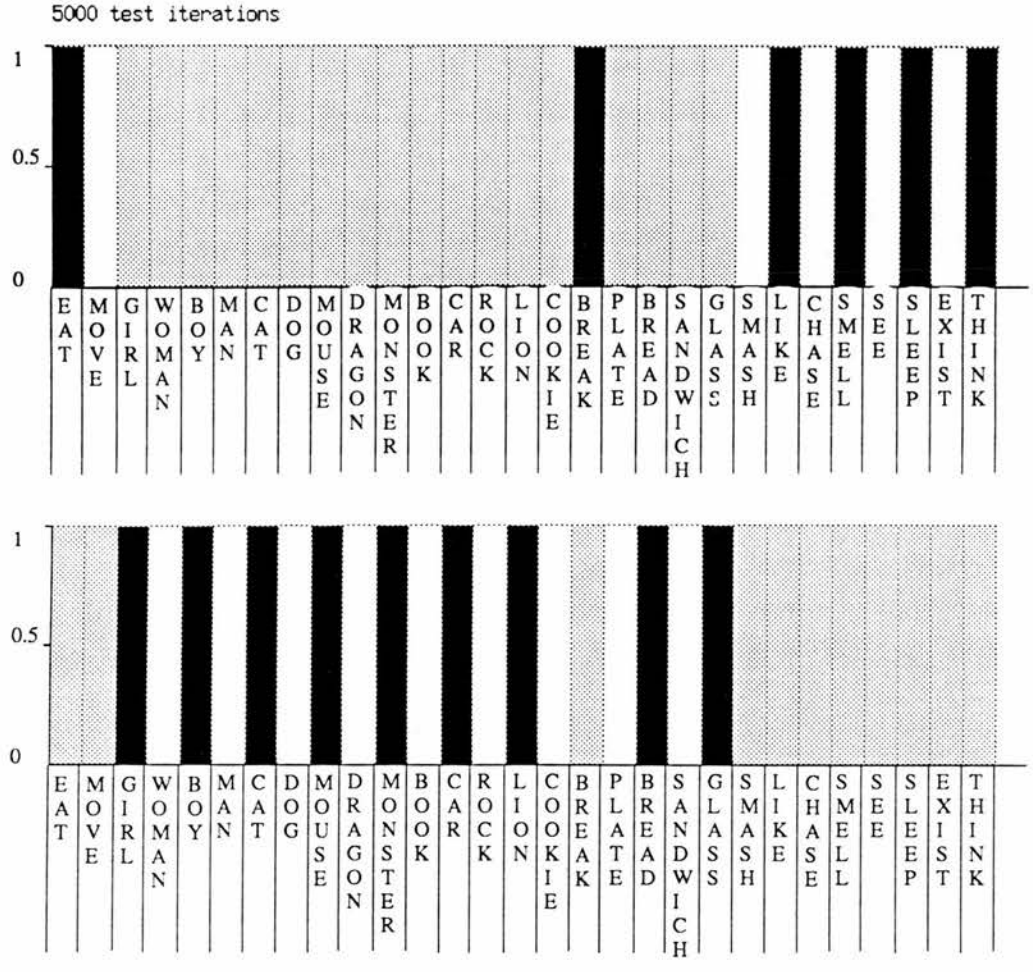
As table 7.11 indicates, 5 of the original categories in the corpus were recovered by output units which responded to the relevant words exclusively. Subsequent investigation revealed that this could be improved to 6 categories if the number of training iterations was increased from 27,354 to 50,000 (unit 3 responds exclusively to 'dragon', 'monster', and 'lion', and thus corresponds to the original category of

NOUN-AGRESS), which is still many fewer than the 136,770 training iterations used by Elman (1988).

Increasing the number of training iterations may perhaps also have improved performance on some of the earlier simulations, but our main intention here is to demonstrate that, even with comparatively small amounts of training, the network is capable of learning to develop useful clusterings of the input data on the basis of its statistical structure. In any case, it is of course important to bear in mind that even perfect performance on the Elman corpus does not guarantee that the network will perform well with the complexities of a real corpus; as with many other areas of interest within the domain of Artificial Intelligence, scaling up from 'toy' problems to real ones may not be straightforward. Failure to perform reasonably well on the Elman corpus, however, would considerably lower our expectations of the network's performance on a real corpus.

Recall that with the parameters used above, the network was able to find interpretable clusters of words from the original corpus which correspond well with those defined by Elman. In the two cluster case, however, a perfect split between nouns and verbs was not quite achieved; two nouns ('plate' and 'glass') were assigned to the group containing verbs. It is of interest to see whether the network could improve upon its performance with this particular syntactic distinction and separate nouns from verbs perfectly. One parameter which can be changed here is the number of window positions around the target word which the network 'sees'. In the above simulations, this number was set at 2. However, what happens if we increase it slightly? Since the sentences in the corpus are short (2 or 3 words in length), we would risk introducing unhelpful noise if the window is increased very much. An increase to 3 positions either side of the target word was, however, explored and it was discovered that with the extra information this provides to the network, perfect performance was indeed possible; the position after 27,534 training iterations and 5,000 test iterations is shown below in figure 7.17.

Figure 7.17: Output Unit Responses After 5,000 Test Iterations



7.8 Conclusions

As we have already noted, it is not of particular interest to seek further improvements in the other analyses considered in this chapter. We have established that, with a corpus whose structure we understand from the outset, the neural network introduced in this chapter can perform as intended, discovering many of the categories present in the corpus using statistical structure as its source of information. Furthermore, its performance compares favourably with that of Elman's much more complicated supervised network.

The next stage in examining the extent to which statistical structure can be useful in categorizing words on the basis of meaning is to provide the network with input taken from a natural language corpus.

8. *APPLYING THE NEURAL NETWORK APPROACH TO A REAL CORPUS*

In Chapter 7 we saw that an unsupervised neural network could learn to categorize Elman's (1988) artificial corpus into syntactic groupings by using the statistical structure inherent in the language input alone.

The question now arises as to whether we can extend this approach to consider a real corpus such as the Wall Street Journal corpus, which was earlier investigated using 'standard' statistical approaches in Chapters 4 and 5.

If we are to do this, we need to be able to deal with the fact that words are likely to be assigned to more than one output cluster according to the contexts in which they are used. Of course, the fact that the network will do this is a very desirable feature and was one of the main reasons for developing it in the first place. However, whilst considering the Elman corpus, this was rarely an issue because lexical ambiguity was not nearly as commonplace there as it would be in a real corpus. In a real corpus, however, the probability of a particular output cluster responding to a particular word during testing is likely no longer to be 0 or 1, but some intermediate value. Numerous clusters may be expected to be involved since polysemy is a pervasive feature of the English language, as any dictionary will confirm.

Given this, we have to consider how we can compare output clusters when these can respond to words with a probability which is between 1 and 0; that is, when a word is no longer assigned to just one of the available clusters in a Boolean fashion, but is distributed probabalistically over the clusters. Of course, in the situation where each target word *is* only assigned to a single output cluster with a probability of 1 (and thus with a probability of 0 to all other clusters), one can attempt to look at the words assigned to each cluster to ascertain whether they are reasonably similar in meaning. This is essentially the approach that was taken in the analyses presented in Chapters 4 and 5. However, in the situation where each output cluster responds to the target

words with different probabilities, it is not clear that this will be a successful approach.

To provide a solution to this difficulty, one possible alternative means of analysis is to examine the way in which a particular target word is distributed over the available output clusters, and then to compare this distribution with that for other words. The assumption behind this is that if the neural network is genuinely clustering the words in the corpus in a coherent fashion, then words which are truly similar in meaning (similarly distributed in the corpus with respect to context) ought to be distributed in a similar manner over the output clusters of the network. Working in reverse, we might then expect to find that words with similar distributions are similar in meaning. Of course, in the more artificial situation when the probability of an output unit responding to the words *is* either 1 or 0, this amounts to saying that words in a particular cluster should be similar in meaning - which is a standard assumption when performing clustering procedures. Such an analysis would thus be generally applicable, being appropriate in cases where ambiguity does not occur as well as those in which it is present.

If words are to be assessed for similarity on the basis of their probability distributions over the output units, a metric permitting the similarity of these distributions to be quantified must be found. One possible measure is the information-theoretic measure of Kullback-Leibler distance (or I-divergence), as used by Pereira, Tishby, and Lee (1993), discussed below (for an introduction to this and other information-theoretic measures, see Plumbley (1991)). However, this will not work straightforwardly in the present situation because of the possible presence of zero probabilities in either one of the distributions being compared[21]. Given this potential difficulty, and in the light of its success when used for the analyses in Chapter 4, it was decided that the Spearman rank correlation coefficient would again be an appropriate metric for comparing the distributions.

[21] Pereira et al. did not enounter this problem using Kullback-Leibler distance because of the characteristics of the clustering technique they used.

This type of approach has something of a precedent in the work of Pereira et al. (1993). These authors were interested in the assignment of words to 'sense classes'. For each word encountered by their system, each sense class would have an associated probability of the word being assigned to that class. However, Pereira et al. restricted their attention to the consideration of words participating in a particular syntactic relationship. This was the relationship between a transitive verb and the head noun of its direct object. In order to find words of the required type, it was first necessary for their corpus to be parsed. By analysing the resulting pairs of verbs and nouns, the distribution of probabilities that each verb would have occurred, given that a particular noun had formed its direct object, could be calculated for each noun. The intention was then to classify the nouns into classes on the basis of these probability distributions.

That this approach is similar in principle to the problem faced in analysing the performance of our unsupervised neural network in dealing with word sense ambiguity is made clear by Pereira et al. :

"In general, we are interested in how to organize a set of linguistic objects such as words according to the contexts in which they occur, for instance grammatical constructions or *n*-grams ... the theoretical analysis outlined here applies to that more general problem, but for now we will only address the more specific problem in which the objects are nouns and the contexts are verbs that take the nouns as direct objects (p184)."

In the present case, for every target word, we are looking at the distribution of probabilities that each output unit will respond, given its occurrence. We wish to obtain such a distribution for each target word being considered, and then, as indicated above, to compare the similarity between these distributions. Pereira et al. did not in fact proceed in this fashion, but compared their probability distributions with the cluster centroids (averaged probability distributions) of clusters which they later generated.

To consider our problem more formally, we have a set of target words w_i and a set of output units λ_j . For each target word, we can establish the probability (at any point during testing) that each output unit will respond to that target word, given that it has

occurred as an input to the network. We can denote this probability for a particular output unit λ_a and a particular target word w_b as:

$$p(\lambda_a \mid w_b), \text{ where } a \in j, b \in i \qquad (8.1),$$

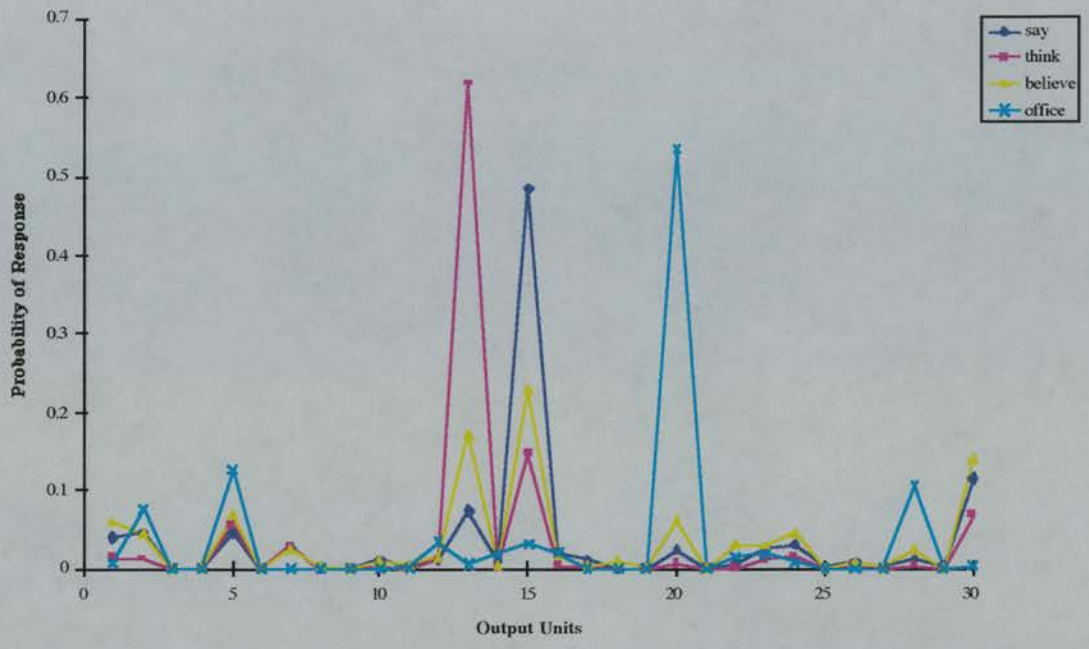
noticing that

$$\sum_{a=1}^{j} p(\lambda_a \mid w_b) = 1 \qquad (8.2).$$

To compare the distribution in 8.2 between different target words w_b, we can treat the j probabilities $p(\lambda_j \mid w_b)$ as components of a vector $\mathbf{w}_b$ and compute the similarity between the vectors. As indicated above, we shall use the Spearman rank correlation coefficient for this purpose. Using this measure, a higher correlation will be accorded to probability distributions which are more similar.

This procedure is illustrated in figure 8.1 below, using some example data from one of the analyses to be considered in more detail later. Four words are considered here: 'say', 'think', 'believe' and 'office'. The analysis revealed that whilst 'think' showed a correlation of 0.943 with the word 'say' over the 30 output units in the neural network used, and the word 'believe' showed a correlation of 0.908 with 'say', the word 'office' showed a lower correlation of 0.699 with 'say'. The probability distributions for the three words are plotted in figure 8.1, and inspection of this makes it possible to some extent to identify the similarities between the distributions for 'say', 'think', and 'believe'.

Figure 8.1: Output Unit Probability Distributions for Four Words in the Wall Street Journal Corpus



We can use this distributional similarity measure as a distance metric for all of the 1000 target words considered. Using this metric, a cluster analysis procedure was carried out in the same way as for the analyses presented in Chapter 4. The resulting dendrogram and table of nearest neighbours[22] are illustrated in Appendix C, figure C.1. Whilst, as we noted in Chapter 6, we wish to avoid heavy reliance on a purely descriptive procedure such as cluster analysis, the procedure is useful for illustrating the relationships between the words considered. It also enables us to confirm that distributional similarity over the output units of the network, which reflects the contextual variation of the words, can produce structures which are intuitively reasonable and familiar. This in turn supports the claim that, even though the Wall Street Journal corpus is vastly more complicated than the corpus considered in Chapter 7, the neural network can perform in the intended fashion. It should be emphasised that, whilst the dendrogram in figure C.1 is superficially similar to those

[22] The same 50 randomly chosen target words that were used in the tables of nearest neighbours for Chapter 4 were also intended for presentation in table C.1. A small number of these target words, however, do not appear in the table because of minor differences between the samples of the corpus used by the neural network method and the earlier vector methods.

shown in Appendix A, the method by which it was obtained is, of course, very different.

In figure C.1, semantically related groups of words are identifiable, despite the increased complexity of the methods used in this chapter over those used in Chapter 4.

Groupings of words which were found to be robust in the earlier analyses also make an appearance here; groups containing the names of individuals ('richard', 'michael', 'robert', 'paul', etc.), numbers, units of time ('year', 'week', 'quarter', 'month', etc.), and nationalities ('british', 'japanese', 'soviet', etc.), for example, are all evident, among numerous others.

As with the earlier analyses, syntactic groupings of words are still evident in table C.1 and figure C.1, with different forms of the same verb sometimes appearing next to each other('do', 'did', 'does', etc.), and the modal verbs grouping together ('could', 'can', 'should', etc.).

In addition to these and other examples of words which are related in meaning being placed together, antonyms are again present, as they were in the earlier analyses carried out. In figure C.1, for example, it is possible to identify several such groupings ('raise' and 'reduce', 'rise' and 'decline', 'higher' and 'lower' etc.), and in table C.1, we find a similar example ('it's' and 'isn't'). As we noted earlier, it is possible to regard antonyms as being very closely related in meaning, and the neural network has done so on the basis of the words' probability distributions over the output units.

8.1 Evaluation

As noted above, the evaluation to be used for the neural network analyses carried out on real corpora is one which involves examining the similarity between words on the basis of the way they are distributed over the available output units.

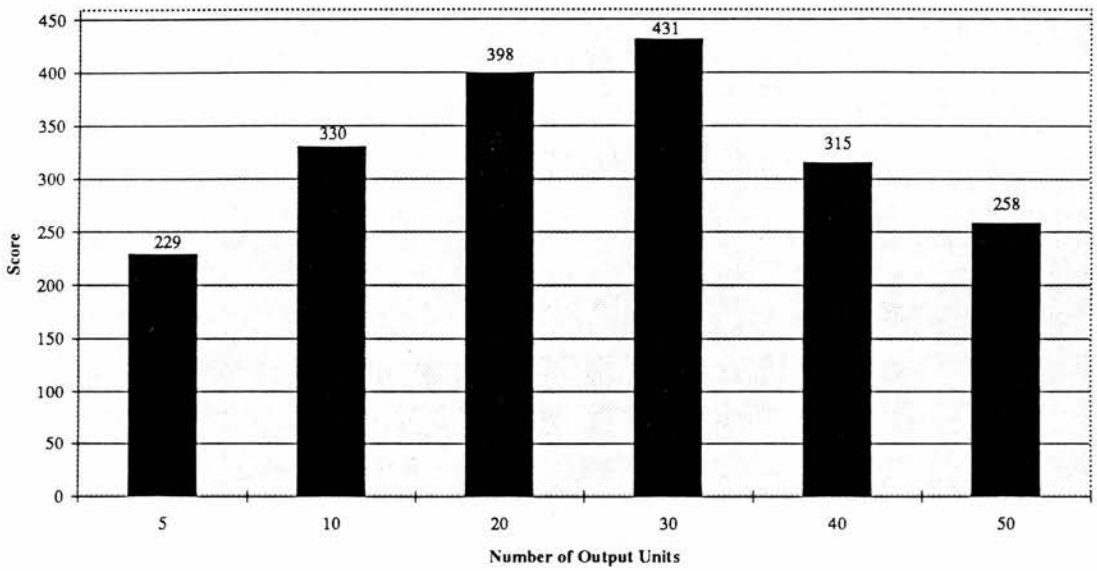
To provide a benchmark here, the method of assessment using Roget's Thesaurus which was described in Chapter 5 was again used. Once a matrix of distributional similarities has been calculated for all pairs of words (using the Spearman correlation coefficient), we can list for each target word concerned the words which are most similar in terms of their distribution over the output units, in order of decreasing similarity. In other words, we can produce an ordered list which gives the nearest neighbours for each of our target words in 'distributional similarity space'. Once we have done this, we can compare the list of 10 neighbours produced for each target word with the appropriate categories for that target word in Roget's Thesaurus, and can determine how closely the two correspond. The matching procedure is described in detail in Chapter 5.

As with all complex systems such as neural networks, we are faced with such a large number of combinations of parameters which could be varied that we cannot realistically explore all of these. In the analyses to be described, attention was focussed on the effect of varying the number of available output units (clusters) in the network on the way in which target words were distributed over them. In particular, it was of interest to see how varying this parameter affected the words' distributional similarities, and consequently to see how it affected their match with the categories contained in Roget's Thesaurus.

The results are presented in figure 8.2. For each of the analyses conducted, the other parameters controlling the network's behaviour were held constant. The learning rate for the winning unit during training was set at 10, and for other output units was set at 1. The window length extended 2 words either side of the target word, this length being chosen because of its superior performance in the work carried out in Chapter

4. The analyses were carried out for 1000 target words (the 1000 most frequent words in the corpus), and 200 context words (the 200 most frequent words in the Wall Street Journal corpus) were considered. For each analysis, 3,000,000 training iterations were carried out (using the Wall Street Journal corpus), followed by 3,000,000 test iterations over which the response probabilities for each of the output units was calculated.

Figure 8.2: Results of Evaluation Using Roget's Thesaurus as a Benchmark



It was important to ensure that, within each of the analyses in figure 8.2, the 10 neighbours considered for each of the target words were reasonably different from those listed for other target words. This is a means of checking whether a sensible clustering had been carried out, before testing for significant differences between the analyses shown in figure 8.2. This check is of exactly the same type as that described in Chapter 5, when ths 'standard' analyses were evaluated against Roget's Thesaurus. The method is explained in detail there, but it is essentially one of checking that those neighbours which contribute towards the score for each target word do not in general also contribute to those of other target words chosen at random and matched in frequency (since pairs of randomly chosen target words would not in general be expected to be similarly distributed over the output units). This is achieved by dividing the 1000 target words into two equally sized groups and calculating the number of

scoring neighbours that are shared between pairs of frequency-matched target words. The results of the check are shown below in table 8.1. For each type of network used, the number of scoring neighbours in each of the 2 groups is given, along with the number of any of those neighbours shared between a pair of target words.

Table 8.1: Number of Neighbours Shared by Randomly Chosen, Frequency-Matched Target Words

Number of Output Units in Network	Number of scoring neighbours in first group	Number of scoring neighbours in second group	Number of 'shared scoring neighbours'
5	111	118	0
10	166	164	0
20	212	186	0
30	223	208	0
40	170	145	0
50	132	126	0

Table 8.1 confirms that the neighbours surrounding the target words in each analysis were indeed different in each case and we can now proceed with confidence to consider the results indicated in figure 8.2. As figure 8.2 indicates, the matching between the groups of words obtained on the basis of distributional similarity over the output units of the network and the groups in Roget's Thesaurus increases up to a point (with a network which has 30 output units), and then begins to decrease. The scores for the 1000 target words were compared between these 6 conditions and an overall significant effect for the number of output units used was found using a Kruskal-Wallis one-way analysis of variance (X^2=36.3392, $p<0.01$). To carry out comparisons between the individual analyses, a Wilcoxon signed ranks test was used, with the resulting statistic converted to a z score, and with an appropriate correction for carrying out multiple comparisons (Siegel and Castellan, 1988). This revealed that the network with 30 output units was a significantly closer match with Roget's Thesarurus than the one with 40 units ($z=-4.6836$, $p<0.05$, 2-tailed). No significant difference was found between networks having 20 units and 30 units, or between neworks having 10 and 20 units. However, a significant difference was obtained between the network having 5 units and the network having 10 units ($z=-3.8004$, $p<0.05$, 2-tailed).

In summary, then, networks having between 10 and 30 output units produced results in closer agreement with Roget's Thesaurus than networks with fewer or more output units. In terms of absolute performance, the network which had 30 output units produced the highest score. It should be noted that all of the scores indicated above are low when compared with the maximum possible score of 10,000. However, as was stressed in Chapter 5, the method of assessment used here is most appropriately regarded as a measure of *relative* performance between a number of different approaches, rather than as an indication of *absolute* performance.

Whether the results obtained are due to some fundamental limitation in the English language on the ways in which words can be polysemous, or whether they are due to the increasingly large ratio between the amount of training data available to each output unit as the number of output units increases, is not clear. To explore the latter possibility, it would be necessary to conduct further analyses using this network with a much larger corpus. Nonetheless, there is some evidence from the work carried out to support our working assumption that words that are distributed in similar ways over the output units in the network may also be similar in meaning. This is by no means true for all words; as figure C.1 in Appendix C illustrates, many words that we would regard as similar in meaning were not distributed in particularly similar ways here. Again, a larger training corpus might have changed this situation. The important feature of the work carried out here and in Chapter 7, however, is that we have made the major step of allowing the network to decide for itself how to deal with the senses into which word tokens can be divided. As we noted earlier, many previous research efforts in this area have not attempted to deal with this issue, and those that have did not do so in an on-line fashion as has been the case with the analyses carried out here. The network was given no information about the very complex language data to be encountered, and yet evidence was found that some of our intuitive notions about word meanings were captured by its performance. It should be remembered also that the network used was of a very simple design which used relatively little information about the words encountered. As was noted with the vector approaches considered in Chapters 4 and 5, even restricting the capabilities of an artificial system in this respect

does not mean that it cannot extract some familiar looking structure from the language data. This, again, lends weight to the idea that intralinguistic information could make a potentially important contribution to the development of a classification of word meanings and senses.

We noted earlier that there is indeed some evidence to support our supposition that words which are similar in meaning in the corpus do have more similar distributions over the output units of the network than words that are more dissimilar in meaning. We may also enquire, however, whether the ways in which target words are distributed over the output units can be related to other measures of the characteristics of those words. In particular, it would be of interest to relate the distributions, which are derived from the statistical structure of the corpus to which the network is exposed, to a *psychological* measure of the differences between words. After all, it is a psychological phenomenon that we are investigating and seeking to account for.

To explore this possibility, the psychological classification of words devised by Jones (1985) was considered. As was discussed in Chapter 5, Jones described a metric known as 'Ease of Predication'. This was operationalized in terms of the ease with which words can be placed into simple factual statements. It was suggested that there might be a semantic factor underlying the correlation between the ease with which a word gives rise to mental imagery, and the ease with which it can be read. Jones was able to show that there was indeed a high correlation between the Ease of Predication of a word and the ease with which an image of it could be formed, noting that:

> "the ease of predication measure therefore provides ... evidence in favor of the frequently voiced hypothesis that apparent effects of imageability may be mediated via a previously unspecified semantically defined variable with which it is closely correlated (p7)".

This measure is, of course, of interest here because the measures resulting from Jones' experiments are the product of human intuitive judgements about word usage. Jones envisaged the Ease of Predication measure as a possible reflection of the differences between the distributions of predicates for various words. Here, however, we have avoided the use of representations for words based on the use of predicates or any kind of conventional features. Nonetheless, it seems reasonable to suppose that the

judgements of Jones' subjects might ultimately have been influenced by knowledge of statistical structure which, as we have seen earlier, can be used for categorizing words on the basis of meaning. In particular, the distribution of words over the output units of the network used here, which is a reflection of the contextual variation of the words, may be a useful measure to relate to Ease of Predication.

Jones (1985) obtained mean Ease of Predication scores for 20 words in each of 5 categories. These categories were high-imageability nouns, low-imageability nouns, adjectives, verbs, and function words. He found that there was a significant effect for the category into which the words were placed, with Ease of Predication scores decreasing from high-imageability nouns, which had the highest scores, to function words, which had the lowest. This coincides with the reading abilities of some patients with deep dyslexia, such as patient G.R., who also read nouns most easily, followed by adjectives, then verbs, and finally function words (see Hinton, Plaut, and Shallice (1993) for a discussion of this phenomenon). For the purposes of the present analysis, all of Jones' words which occurred amongst the most frequent 3,000 words in the Wall Street Journal were selected[23]. This resulted in the selection of 15 high-imageability nouns, 16 low-imageability nouns, 16 adjectives, 18 verbs, and 19 function words. For each word, a crude measure of the variability of the contexts in which each word had occurred was also calculated. This measure, which we shall refer to as the 'spread' of a word over the available output units, is simply a count of how many output units are involved in representing that word. If the spread is high, the word is widely distributed over the output units; if it is low, the word is constrained in its distribution over the units.

For practical purposes, it was more straightforward, however, to calculate the number of units *not* involved in representing a word. This was achieved by examining the probability distribution of each word over the output units of the network with 30

[23] It was initially intended to use words occurring amongst only the 1,000 most frequent words in the Wall Street Journal corpus, as in the other analyses presented in this thesis. However, few of Jones' words occurred as frequently as this, and the decision was therefore made to include some less frequent items.

output units[24] and counting the number of units whose probability of responding to the word was zero. In other words, the number of units which played no part in representing each word was counted. If this number is high, the word's 'spread' is low. On the other hand, if this number is low, this corresponds to a high 'spread' value. The words in each category, along with their frequency in the 10,000,000 word Wall Street Journal corpus, and the number of units *not* contributing to the representation of each word are given in table 8.2 below. It should be noted that for each word, its 'spread' is the difference between the total number of output units available (which is 30 in this case), and the number of units not contributing to its representation.

Table 8.2: Ease of Predication Scores, Word Frequency, and Units Contributing

Category of Word	Word	Mean Ease of Predication Score Jones (1985)	Frequency in the Wall Street Journal Corpus	Number of Units Not Contributing to the Representation of the Word.
High-imageability noun	body	6.50	442	8
	child	6.64	555	11
	door	6.57	395	14
	hall	6.14	524	3
	meat	6.86	226	17
	newspaper	6.64	1077	8
	officer	6.43	4481	7
	sea	6.64	422	10
	woman	6.71	696	9
	car	6.71	2341	4
	city	6.79	3541	3
	lake	6.50	332	3
	mountain	6.79	274	11
	river	6.86	498	11
	valley	6.14	586	10
Low-imageability noun	direction	4.71	599	11
	effort	4.14	1827	8
	hour	6.21	554	16
	law	5.36	4475	5
	life	5.07	2726	5
	moment	4.64	471	19
	opinion	5.00	665	9

[24] This particular network was used since the evaluations presented earlier in this chapter suggested that it would produce optimal performance.

Table 8.2 (contd.)

Low-imageability noun (contd.)	thought	5.36	1307	12
	duty	5.21	291	15
	fact	5.36	1858	21
	health	5.00	2519	4
	history	5.93	1177	9
	knowledge	4.79	364	14
	method	5.71	266	16
	month	6.43	5630	15
	truth	4.79	268	18
Adjective	bright	3.71	201	11
	every	2.00	2219	2
	foreign	4.21	5689	6
	fresh	4.00	387	13
	happy	4.43	385	11
	important	3.64	2184	14
	rich	5.00	549	7
	soft	3.86	357	13
	various	3.00	1096	6
	different	3.71	1544	12
	famous	4.64	269	14
	former	3.86	4372	3
	golden	4.00	328	13
	heavy	4.36	1484	7
	simple	4.07	447	9
	sudden	3.36	285	19
Verb	accept	2.93	758	8
	believe	3.29	2043	3
	continue	3.86	2840	9
	eat	3.71	193	16
	grow	3.71	708	11
	include	3.29	2352	1
	lose	3.64	693	8
	prepare	4.71	203	22
	understand	3.86	527	25
	ask	3.86	782	12
	consider	3.50	1472	5
	enjoy	3.64	192	16
	hear	3.64	477	17
	know	3.00	2385	3
	obtain	3.50	410	15
	receive	4.00	1289	7
	sit	4.43	281	17
	write	4.36	451	12
Function Word	him	2.43	4798	6
	it	2.43	64095	3
	you	3.14	7451	9
	at	1.57	54726	3
	in	2.07	205401	5
	that	2.29	101928	5
	what	1.71	8611	3
	who	2.29	19698	9
	has	1.64	38182	5
	I	3.50	11309	5
	she	2.86	5033	1
	and	1.57	201195	5
	but	1.29	38823	6
	out	2.86	11572	7

Table 8.2 (contd.)

Function Word (contd.)	this	2.29	27052	7
	which	1.79	26499	5
	did	1.79	2647	3
	is	1.43	81085	6
	would	1.64	25349	9

As we have already noted, our intention here is to examine the relationship between Ease of Predication and the contextual variation of words over the output units of the neural network. Our prediction regarding this relationship is that words whose Ease of Predication score is high ought to be more constrained in their 'spread' over the output units than those whose Ease of Predication score is low. In other words, we are predicting that a high Ease of Predication score would tend to indicate a relatively small amount of contextual variation.

At first sight this might seem a surprising prediction, since Ease of Predication was operationalized by Jones as the ease with which words can be placed into factual statements; surely it might be argued that those words which vary widely with respect to the contexts in which they occur should be easier to place into such statements than words which show little such variation? The reason why this is *not* the prediction we are making is because of the behaviour of closed class words. This category of word, which Jones identified as having the lowest Ease of Predication scores of all the categories he considered, was also noted during the modelling being described to be highly variable with respect to context. This suggested that wide variability of context may not guarantee that the task of placing such words into factual statements will be easy.

To test the prediction, a stepwise linear regression analysis was carried out using the data shown in table 8.2. In this analysis, the spread of a word over the output units of the network and word frequency in the Wall Street Journal corpus were both considered to examine their relationship to the dependent variable - Ease of Predication. Word frequency was included because, as table 8.2 makes clear, there were very large differences in frequency between the words involved. In particular, the function words had very much higher frequencies than most other types of words.

The influence of this sort of effect clearly needs to be assessed and taken into consideration.

The linear regression revealed that the network 'spread' measure and word frequency together accounted for a significant amount of the variance in the Ease of Predication measure (R^2=0.234, F(2,81)=13.664, p<0.001). The analysis also revealed a significant positive partial regression coefficient between the spread measure and Ease of Predication (β=0.229, p<0.02, 1 tailed), which indicates that Ease of Predication increases as the number of units used to represent the words decreases (and therefore as the number of units *not* used increases). Furthermore, there was a significant negative partial regression coefficient between word frequency and Ease of Predication (β=-0.388, p<0.001, 1 tailed), showing that Ease of Predication increases as word frequency decreases.

These findings support the prediction that higher Ease of Predication scores are associated with more restricted spread over the output units of the network, which in turn reflects reduced contextual variation. However, they do also reveal that word frequency is also closely related to the Ease of Predication scores. We can thus support the notion that Jones' (1985) Ease of Predication scores can be predicted to a significant extent by the network 'spread' measure; those words which have the lowest Ease of Predication scores are those which exhibit the most contextual variation, whilst those which have higher Ease of Predication scores tend to show less contextual variation. The concordance here between our measure of the network's performance and Jones' psychological measure is of interest because it suggests that the two measures may both be a reflection of a single underlying phenomenon.

This underlying factor may be that of the influence of statistical structure. We know that the network is governed by the statistical behaviour of the language data, since this is the only information with which it is provided, yet it distributes target words over the output units of the network in a way which is predictive of Ease of Predication, a measure that is the product of the intuition of Jones' subjects. As we noted earlier, these findings may be an indication that the responses of Jones' subjects

reflect their knowledge of the English language and the statistical behaviour of words and their contexts within it. If so, these results provide supplementary experimental support to the empirical computational work carried out earlier in this thesis. We have already seen that the statistical structure of language can be of considerable use in developing a categorization for word meanings, and now we have an indication that human subjects' ability to judge the ease with which words can be placed into factual statements may be influenced by knowledge of statistical structure.

It is important to bear in mind, nonetheless, that word frequency is also significantly negatively correlated with the Ease of Predication measure. This finding is perhaps not surprising since function words, which are assigned the lowest Ease of Predication scores, are among the most frequent words in the Wall Street Journal corpus. Thus, a significant part of the scores obtained by Jones can be predicted on the basis of the frequencies of the words involved.

We must also note that, whilst word frequency and the network spread measure both play a significant part in explaining the variance present in the Ease of Predication measure, they are themselves correlated (Pearson r=-0.280, $p<0.01$, 1 tailed). This indicates that more frequent words are significantly more spread out over the output units of the network than less frequent ones, and suggests that contextual variation and word frequency are not independent of one another despite the fact that the neural network used was designed to minimize the effects of word frequency when learning about context.

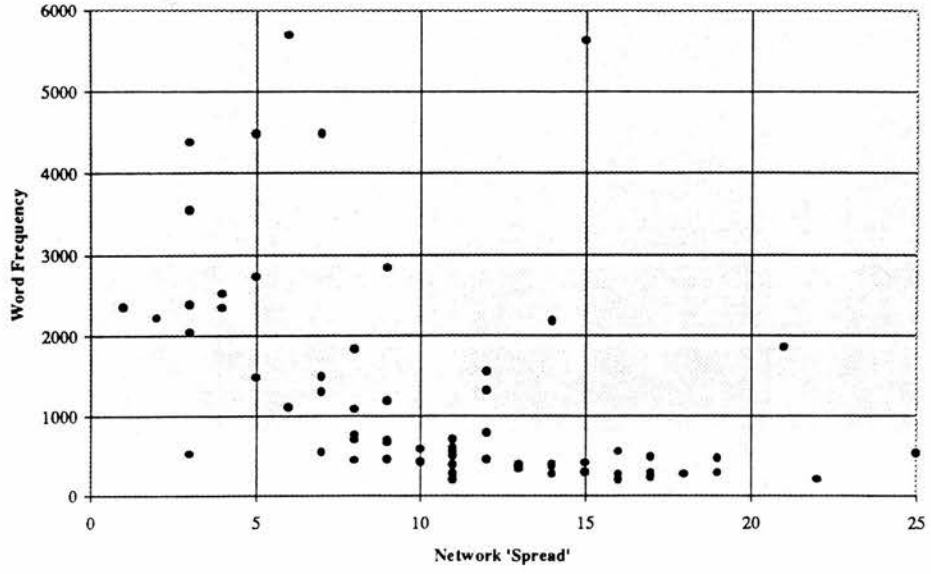
At first sight, this is a surprising finding. As was described in Chapter 7, to prevent word frequency from distorting the way in which the neural network dealt with the target words, it was arranged that learning for each word would decrease progressively on each occasion that it was encountered. The amount of learning for frequent words would thus rapidly decrease, whilst remaining relatively large for less frequent words. Why then should the frequency of a word have any relationship to its contextual variation?

The answer here perhaps lies in the fact that, whilst the network was designed not to allow a word's frequency to influence its behaviour, guaranteed avoidance of such an effect here does assume the provision of an infinite amount of language data. This is because, in the limit, the differences in the amount of learning applied to each word would be negligible. However, we do not of course have an infinite amount of data with the large corpus used as input for the network in this chapter, nor even with the large amount of language data encountered by human beings. Some words will still occur more frequently than others, and, in practice, the word frequency effect on learning may still be present. In particular, those words which occur more frequently will have had the opportunity to be encountered in a wider variety of contexts than less frequent words, precisely because they are more frequent. In this way, word frequency and contextual variation are confounded.

A correlation between word frequency and ambiguity has been noted before in natural language research. Miller (1963), for example, discusses empirical work showing that nouns, verbs and adjectives with the highest frequencies tend also to have the greatest number of definitions. If we exclude the function words, so that we, too, only consider nouns, verbs, and adjectives, we do indeed find a significant correlation between word frequency and the network 'spread' measure[25] (Pearson r=-0.482, $p<0.0005$, 1 tailed). In this case, of course, the correlation is negative because our measure of the network's performance counts units which are *not* used in representing each word. Furthermore, the correlation is larger than when we included function words earlier. This difference is due to the effect of the very large increase in frequencies for function words, which is now no longer present. The data are plotted below in figure 8.3.

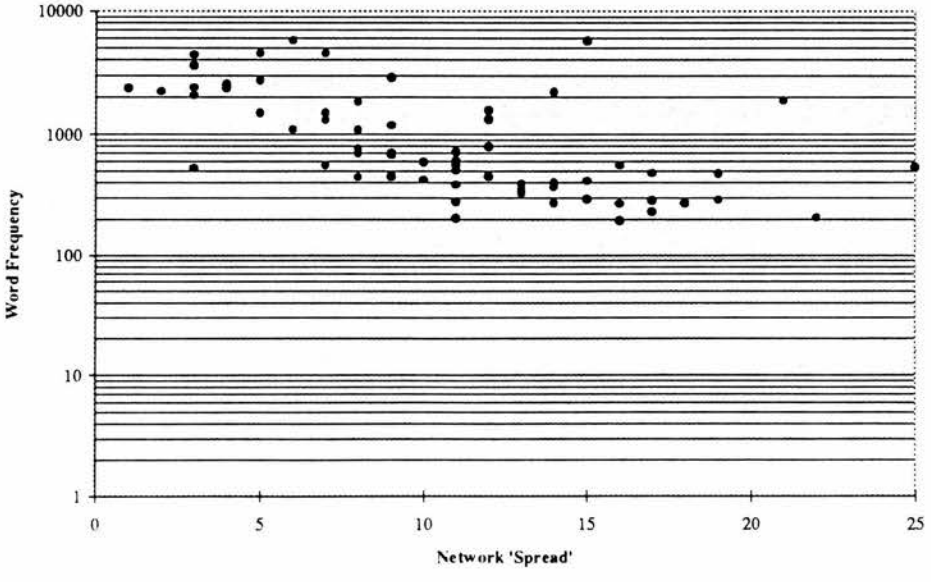
[25] An experiment is also being conducted with Peter Hipwell of the Centre for Cognitive Science, The University of Edinburgh, to determine whether a correlation also obtains between lexical decision latencies and the network spread measure.

Figure 8.3: Correlation Between Network 'Spread' and Word Frequency



To make the relationship between the two variables easier to appreciate, the data are also presented in figure 8.4, where the axis representing word frequency is drawn logarithmically.

Figure 8.4: Correlation Between Network 'Spread' and Word Frequency



When we use the contextual variation exhibited in the network spread measure, then, we find an even stronger relationship between spread and word frequency if we exclude the function words. This is in close agreement with the findings reported by Miller (1963).

Psycholinguistics has, however, often taken the view that the effects of word frequency and context are independent of one another. To take a historically influential example within Psycholinguistics, Morton (1970) has proposed a model of these effects upon word recognition. In representing words as discrete entities, and using an activation metaphor, this model prefigures most later modelling of word recognition. In Morton's approach, it is assumed that each of the words an individual knows has a corresponding 'logogen'. Each logogen has a threshold, and if the perceptual input for a word gives rise to a level of activation which exceeds this threshold, the logogen will fire. At this point, the word is recognized and all logogens return to their resting level of activation. The experimental finding that recognition of frequent words is more rapid than for infrequent words is explained in the logogen model by assuming that the logogens for frequent words have lower thresholds than those for infrequent words. Morton suggested that this would be achieved by reducing the threshold of a logogen slightly each time the corresponding word is encountered. As he points out:

"In the language of Signal Detection Theory, the Logogen Model would predict that the word-frequency effect is due to a difference in criteria between high- and low- frequency words and not a difference in sensitivity (p207)."

In addition to perceptual input, information about the context in which words occur is also instrumental in increasing the activation for a logogen in Morton's model. As a consequence, less perceptual information may be required to recognize a word in context.

Morton's model is concerned with word recognition rather than the issue of word sense disambiguation. It does, however, illustrate that context and word frequency are typically regarded as independent phenomena, with word frequency making its effect through a straightforward Hebbian process rather than through any association with contextual variation.

Tabossi and Zardon (1993) have provided a useful review of studies of the influences of word frequency and context in lexical access, as opposed to word recognition. Again, these influences are typically regarded as being separate from one another. Tabossi and Zardon note that three major classes of theory can be identified. Firstly, exhaustive theories propose that all meanings of ambiguous words are initially accessed in parallel, regardless of any biasing context or of differences in frequency. These studies are supported by cross-modal priming experiments (see, for example, Onifer and Swinney (1981)). The second type of theory assumes that an ordered search takes place, with the meanings of an ambiguous word being searched in order of frequency. When a match occurs, the search terminates, whilst a failure to match causes the search to be extended to the next most frequent meaning. Hogaboam and Perfetti (1975) defended this account following experiments which revealed that subjects were faster to detect an ambiguity when a context biased the less frequent meaning of an ambiguous word than when it biased the more frequent meaning; this was taken to suggest that both meanings had been accessed in the former case, but only one on the latter case. The third class of theory discussed by Tabossi and Zardon (1993) is that which regards lexical access as being sensitive to contextual information and not to word frequency; cross-modal priming experiments carried out by Simpson (1981), for example, have suggested that, in the presence of strongly biasing context, only the contextually relevant meaning of an ambiguous word is accessed.

Tabossi and Zardon themselves carried out cross-modal priming experiments which suggested that, with context which strongly biased the most frequent meaning of an ambiguous word, only that meaning was accessed. However, if the context strongly biased the less frequent meaning of the word, both this and the less frequent meaning would be accessed. These findings thus indicate an interaction between word frequency and context effects, and were interpreted in terms of the time course of activation of an ambiguous word. In isolation, these authors suggest, the most frequent meaning of an ambiguous word will have a stronger, faster, and longer lasting activation. In the presence of context which biases a particular meaning, however, the activation of that meaning is strengthened and speeded up at the expense of the other.

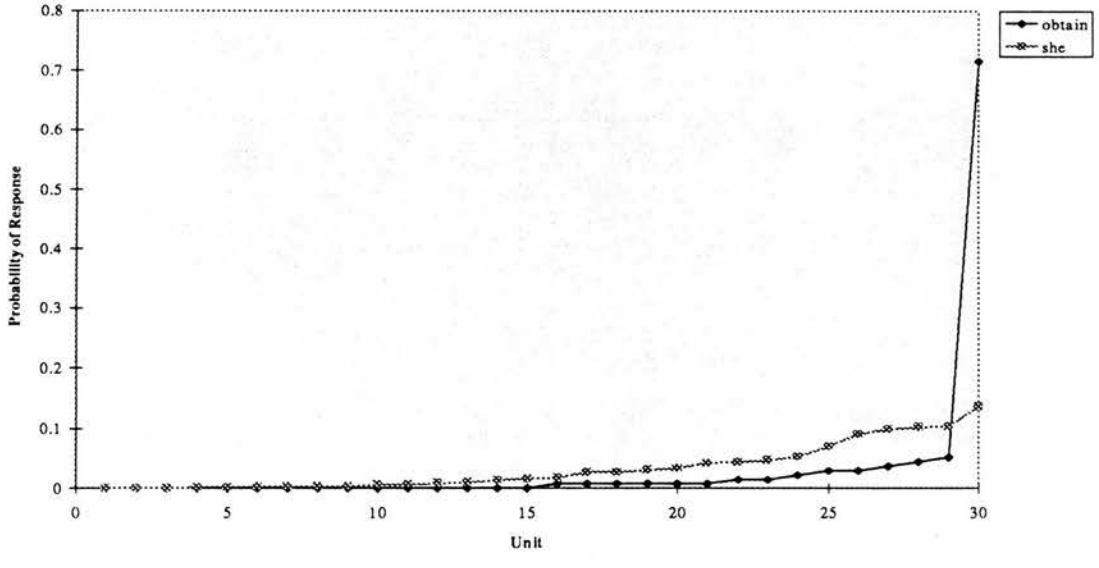
In the light of our analysis of the performance of the neural network described above, the independence of word frequency and context effects appears to be open to question. In particular, we have observed empirically that more frequent words will tend to be encountered in a wider variety of contexts than less frequent words (although word frequency and contextual variation do also play separate, significant roles in predicting Ease of Predication). Since the set of contexts in which a high frequency word may be placed is likely to be larger than that for a low frequency word, we might predict that context will be more important in determining the appropriate sense for high frequency words such as, for example, 'get', 'make', and 'eat'. Equally, we might state that the psychological status of high frequency words is in large part due to their being encoded and defined in terms of their contexts. For instance, in the standard visual lexical decision task, high frequency words are typically recognized faster than low frequency words. If the task were construed as "recall or generate a context legitimately containing the target word", then the high contextual variation of high frequency words would correctly predict a fast response.

It might be objected that this does not follow, however, because it assumes that the probability distribution of contexts is not markedly different between different words. However, if, for example, the large set of contexts for a high frequency word is dominated by a small number of contexts which have a high probability of occurrence, and a large number of contexts whose probability of occurrence is negligible, then correct prediction of the intended sense of the word would not be difficult on average. It might, indeed, be easier than predicting the intended sense of a low frequency word with contexts whose probability of occurrence is more similar, even though the set of contexts themselves is small. However, in practice, the probability distributions for the contexts of words appear unlikely to differ in such a way. This is because, as Zipf (1945) has noted, the frequency of the different definitions for a word follows a similar distribution to that which obtains between a word's frequency and its rank (which we discussed earlier in Chapter 3). We might then expect that, whilst the set of contexts may grow larger with a word's frequency, the shape of the probability distribution for the set will be a similar shape for each word. This of course

presupposes that the contextual variation observed over the output units of the network and Zipf's sets of 'meanings' are related phenomena.

We can examine this by plotting the probability distributions over the output units of the neural network for different words. In figure 8.5 below, the distributions have been plotted in order of increasing frequency[26] for the words 'obtain' and 'she'. These words were chosen because they have very different 'spread' measures - 'obtain' has a value of 15, whilst 'she' has a value of 1. This implies, of course, that 'she' is a word with greater contextual variation than 'obtain' and that the set of contexts for 'she' is therefore larger than for 'obtain'. Nonetheless, it is clear that there is some evidence of the sort of relationship noted by Zipf (1945). This is particularly conspicuous for the word 'obtain', where the plot does not follow a linear path, but one which has a rapidly increasing gradient.

Figure 8.5: Ordered Probability Distributions for Two Words in the Wall Street Journal Corpus

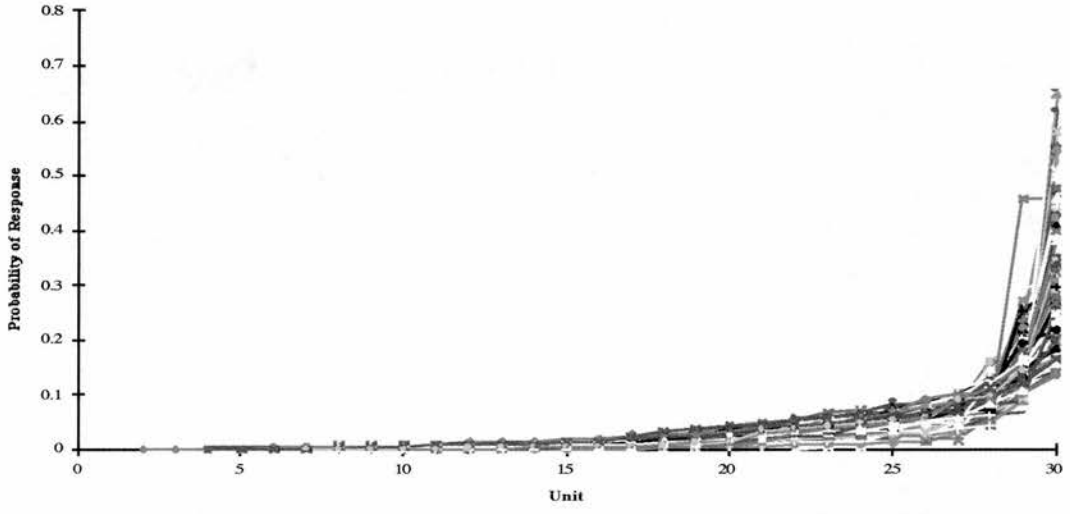


The way in which the network distributes these two target words probabalistically over the output units is reminiscent, then, of Zipf's observations about the frequency distributions of the definitions of words. It is of interest to examine whether this is true for all of the target words we have been considering. Rather than plotting the

[26] Note that, since we are plotting units in order of increasing frequency for each word, the *specific* unit to which the corresponding point on each distribution refers will not necessarily be the same.

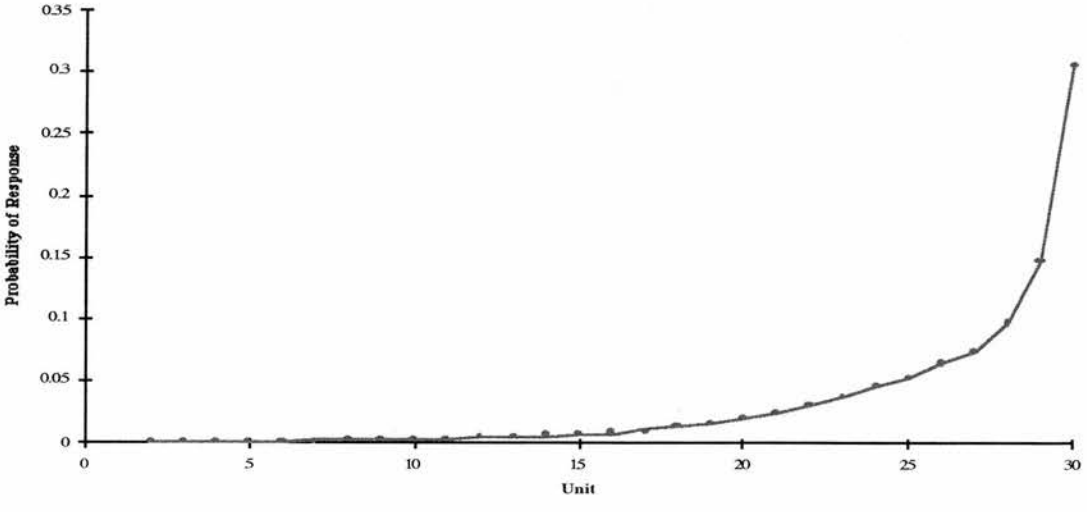
distribution for each separately, we can superimpose all the plots on the same set of axes, as shown in figure 8.6. Any major discrepancies from the relationship we are expecting should then be clearly exposed.

Figure 8.6: Ordered Probability Distributions for Words in the Wall Street Journal Corpus (Superimposed)



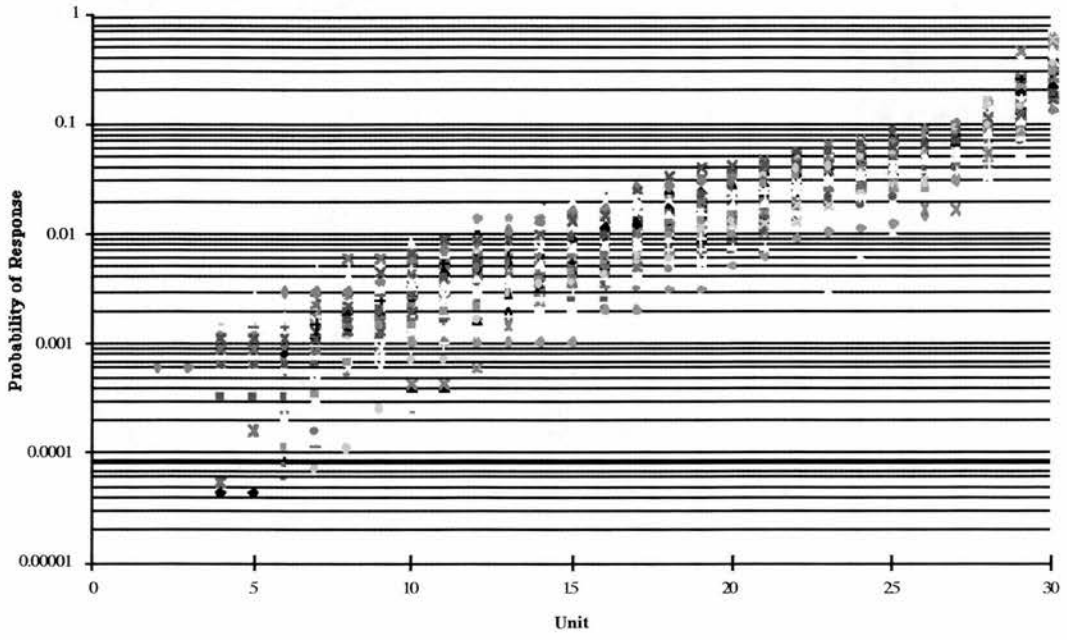
It is clear from figure 8.6 that when the ordered probability distributions for all of the 84 words we have considered are superimposed, we do indeed see evidence of the sort of equilateral hyperbola noted by Zipf. The data can also be summarised by plotting the median probability for the units in each of the ordered distributions, as shown in figure 8.7. This, as we should expect, gives the same characteristic curve.

Figure 8.7: Distribution of Median Probabilities



The data in figure 8.6 can also be depicted with the probability of response being plotted logarithmically, as shown in figure 8.8. It is again clear that, as we would predict, the general trend is for linear relationships between the units placed in rank order and the log probability of response for each unit.

Figure 8.8: Ordered Probability Distributions for Words in the Wall Street Journal Corpus (Superimposed)



What we have seen, then, is that there is evidence that more frequent words occur in a wider variety of contexts than lower frequency words, but that the shape of the probability distribution over the contexts of occurrence is broadly similar between words, regardless of frequency. This shape is typically not a flat distribution or one which shows a linear increase when ordered, but rather one which forms the sort of hyperbola described by Zipf (1945).

If these findings are correct in general, we can maintain our earlier suggestion that high frequency words will, in the absence of any information about their meaning, be more ambiguous than low frequency words. This would, in turn, suggest that the presence of context will be of greater importance in the case of high frequency than low frequency words. The increase in ambiguity for high frequency words appears to be particularly marked for function words; for example, the word 'she' in figure 8.5 is

very spread out over the available output units of the network. Extreme cases such as this may perhaps also account for an anomaly in children's acquisition of language. This anomaly is that, although common words are typically learned before rare ones, the use of function words does not occur until after many content words have been used (Wolff, 1988). Function words are, of course, among the most frequent words in English, and would otherwise be expected to be used early on. The large amount of contextual variation exhibited by these words, however, may make appropriate usage of these words hard to determine until considerable progress has been made in acquiring other aspects of the language.

8.2 Conclusions

In this chapter we have examined the performance of the network introduced in Chapter 7 on the 10,000,000 word Wall Street Journal corpus. A means for comparing words on the basis of their probability distributions over the output units of the neural network was discussed and then applied to 1000 target words in the corpus. We found that, as predicted, words which are intuitively similar in meaning do indeed sometimes show a high degree of similarity in the way they are distributed over the output units. Using distributional similarity as a distance metric, a number of different networks were then compared and evaluated by determining which of them produced groups of words which were most similar to those listed in Roget's Thesaurus. The results suggested that a network with 30 output units produced the best performance here.

In further analyses, we saw that a crude measure of contextual variation over the output units of the network could account for a significant amount of the variance in Jones' (1985) psychological measure of Ease of Predication. Word frequency was also found to account for a significant amount of the variance in Ease of Predication, and was, furthermore, found to be significantly correlated with our measure of contextual variation. The relationship between word frequency and context was then discussed in relation to some of the psychological perspectives on their effects. In particular, we saw that Zipf's (1945) observations about word definitions appear to apply to statistically-defined context; the probability distributions for target words

were seen to show the same characteristic departure from flatness that was noted by Zipf.

The work discussed in this chapter shows that the neural network developed in this thesis has been successful in distinguishing between words, not only on the basis of their statistical context, but more specifically on the basis of that of the senses in which they can be used. This, of course, does satisfy our primary objective of allowing statistical structure to select more than a single representation for each word. Whilst the situation is much less clear cut than it was when using Elman's (1988) corpus in Chapter 7, we have seen that the complexities of dealing with real natural language data are not so great that a simple neural network, which makes few assumptions about the problem, cannot produce encouraging and thought provoking results.

9. *Final Conclusions*

9.1 Further Work

The network we have discussed in Chapters 7 and 8 was developed to allow a categorization of word meanings to be developed which would be a fairer reflection of the potential of the information provided by statistical structure than less sophisticated methods which permit only a single representation for each word. The results presented in Chapter 8 suggested that Jones' (1985) psychological measure of Ease of Predication may be influenced by the statistical behaviour of words, and we also considered the psychological consequences of the relationship between word frequency and contextual variation.

That further psychological implications may arise from an analysis of the network's performance is clearly a possibility. One particular area of enquiry which might fruitfully be explored in the future concerns the effects of damage to a neural network of this kind. Psychologists attempting to account for disorders in which semantic impairments of various kinds are apparent have sometimes done so by examining the consequences of damaging, or 'lesioning', supervised neural networks which learn, for example, to associate orthography with semantic representations.

Hinton and Shallice (1991), for example, explored this approach using a two-layer attractor network. Their motivation was to explain some of the characteristics exhibited by patients suffering from deep dyslexia, but their approach is of particular interest in the present context since it is concerned with the issue of representing word meanings, and because it relates to the findings of Jones (1985) which we discussed in Chapter 8. We shall now briefly consider the main points of their analysis.

Patients with deep dyslexia often make semantic errors when reading a printed word; thus, the word 'peach' might be read as 'apricot'. However, errors of this kind are believed not to be due to a difficulty in selecting the appropriate name for an object. Visual errors are also sometimes present, in which the patient will read a word as another word which is visually similar to it; thus, the word 'patent' might be read as

'patient'. It is generally accepted that in deep dyslexia, the 'phonological route' to word naming is not available, and that the 'semantic route' is used instead; thus, words can be pronounced only through accessing their meaning. Hinton and Shallice proposed that the characteristics of deep dyslexia might be explored by working on the assumption that basins of attraction are present in semantic space. When mapping from orthography to semantics, a system such as a neural network would then be able to map visually similar words to nearby points in semantic space, and then distinguish between the two meanings by causing them to fall into different basins of attraction. The network would thus be able to map the orthography of a word to any location within a particular basin of attraction, rather than to a specific point in the semantic space. Damage in such a situation might consequently give rise to semantic errors or to visual ones. This arrangement is also a desirable one when working with supervised neural networks, which are most easily trained to associate similar inputs with similar outputs.

Hinton and Shallice represented orthography at the input units of their network in such a way that each of the 28 units represented a particular letter of the alphabet and its position within the word in which it occurred. The 68 output units of the network represented the semantics of the words to be examined. Each unit was used to represent the conjunction of a 'role' and its 'filler'. This approach to representing semantics is similar to that incorporated within semantic networks (Collins and Quillian, 1969), in which each word would have a role, such as 'IS AN __', and a filler, such as 'ANIMAL'. For the semantic features to be used by the network, Hinton and Shallice used such features as 'made-of-metal' and 'found-near-sea'. They used a set of 40 words of 3 or 4 letters in length which fell into 5 concrete categories such as 'indoor objects' and 'foods'.

The network developed by Hinton and Shallice, which incorporated a layer of 40 hidden units, was trained using a variant of the backpropagation algorithm (Rumelhart, Hinton, and Williams, 1986). To facilitate the development of attractors, interconnections were arranged between subsets of the output units, and between the output units and a bank of 'clean-up' units. Inspection of the evolution of the

activations of the output units subsequently revealed that attractor states were indeed present.

Hinton et al. used three methods for 'lesioning' the network. The first of these was the removal of a proportion of the connections between a particular pair of layers. The second was the addition of noise to a proportion of the weights, and the final approach considered was to remove a proportion of either the hidden units or the clean-up units. To quantify the effects of the lesions, errors were assessed by calculating the cosine of the angle between the semantic output vector produced by the network for a particular word and that which was the true semantic vector for that word. In general, the results showed that lesions affecting the input to the semantic layer of the network had a greater effect than lesions elsewhere. Furthermore, lesions to the clean-up layer were less disruptive to performance than lesions in other locations. Both semantic errors and visual errors were observed with all types of lesions explored, as well as those involving a mixture of both semantic *and* visual errors, and errors of different types. The incidence of all these types of errors was greater than would be expected by chance. However, the exception to this general finding was that removal of the connections between the output units and the clean-up units produced few errors. A further finding of interest was that when the network produced a response for a word which was relatively far from the correct one, it nonetheless performed at levels above chance when forced to select the appropriate semantic category for the word. This result was obtained by calculating the proximity of the response vector to the centroids of the various categories used.

Hinton, Plaut and Shallice (1993) have reported that the findings of Jones (1985) can be used to produce further behaviour in the network which resembles that seen in some patients with deep dyslexia. Such patients, as we noted in Chapter 8, find nouns easier to read than adjectives, which are in turn easier to read than verbs and function words. Jones suggested that this phenomenon is related to Ease of Predication. When Hinton et al. modelled words by varying the number of semantic features used to represent them, with concrete words being represented by more features than abstract ones, they found that lesions below the level of the clean-up units could produce the

observed effects. Concrete words were found to produce fewer errors because, since they had a less sparse semantic representation than abstract words, this representation contained a greater amount of redundancy. However, damage to the clean-up units themselves reversed the effect, with the network reading concrete words less well than abstract ones, demonstrating the important role of the clean-up units in dealing with the semantic representations for concrete words.

The network used by Hinton and Shallice (1991) differs considerably from the one we have considered in this thesis. Their network is a supervised one, designed to allow the development of attractors, which were considered to be potentially important in explaining the presence of semantic errors. It also makes use of two main layers of units, and represents inputs and outputs in a distributed fashion.

By contrast, the network used here was deliberately designed to be as simple as possible, being trained in an unsupervised fashion and possessing only a single layer of units. However, whilst the two approaches are different, there is an issue which is of relevance to both. This concerns the means by which the semantics of words can be represented. Hinton and Shallice used words placed into 5 categories using a feature-based representation, whilst the network described here used statistical information to decide for itself the relationships between the words it encountered. Analysis of the network used here subsequently showed that such information may underlie psychological phenomena relating to various lexical processes, such as those discussed by Jones (1985). It is worth noting that Plaut and Shallice (1993), in an assessment of the work of Hinton and Shallice, do acknowledge the drawbacks of using feature-based representations:

"... it is not particularly plausible that the semantic representations of a word in the human cognitive system is based on individual feature units at the level of *found-on-farms* and *used-for-recreation*. However, these representations exhibit the characteristics that are essential for demonstrating the influences of both visual and semantic similarity on deep dyslexic reading ... (p388)".

Hinton and Shallice's attractor network exhibited various characteristics observed in deep dyslexia when lesioned. Whilst the network used here does not perform a mapping from orthography to semantics, but is solely concerned with determining similarities and differences between words on the basis of statistical structure,

lesioning it might nonetheless have implications for our understanding of the way in which human beings represent word meanings. Some of the methods used by Hinton and Shallice would seem to be appropriate in doing this. Removing a proportion of the connections between the input and output units would mean that aspects of context, or the target word, or both, would be represented in an impoverished fashion. One interesting avenue of enquiry here would be to examine whether such damage would produce semantic errors of any consistent kind. Since the input representations are sparse and are not of a distributed kind, the expectation would be that errors of this type would *not* occur. Nonetheless, this is an empirical question which might, at least, raise issues about the appropriateness of using localist coding in such a network. The addition of noise to a proportion of the weights in the network would, again, be of interest in assessing the extent to which damage would affect the way in which a word is represented over the output units of the network. With both these types of lesioning, furthermore, it would be intriguing to examine whether the effects of damage are equally severe over all contexts in which the word occurs, or whether they affect those associated with a low or a high probability of occurrence differently. The option of removing entire units, whilst perhaps worthy of investigation, appears at first sight to be rather a crude method of inflicting damage on a network which uses localist coding, and might not therefore be appropriate for the network used here.

The effects of lesioning carried out in these sorts of ways might be assessed in a similar way to the analyses carried out in Chapter 8, with comparision being made between the nearest neighbour assigned to a word in an unlesioned network and that assigned to it in a lesioned architecture. This approach would also allow category effects to be assessed, perhaps by using the benchmark provided by Roget's Thesaurus, as we have done in Chapters 5 and 8. We have observed in Chapter 8 that the probability distribution over the contexts in which words occur is typically very far from being flat, with the bulk of the word's occurrences being assigned to a single output unit. We also have evidence that words of similar meaning share a similar distribution over the output units. The implication of these findings is that damage to the output unit which represents most of the contexts of occurrence for a group of

words could have severe consequences when attempting to understand the meaning of that group of words. We have noted above that simply removing an output unit may be too crude an approach, but it may nonetheless be that there are other forms of damage which can exploit the fact that, statistically speaking, words tend to be strongly biased towards occurring in a particular type of context.

With the network described in this thesis, there is another broad avenue of enquiry which might usefully be explored in the future. This concerns the importance of using both left and right context in representing a word, the issue of the optimum window length for gathering statistical information about target words, and the question of which type of context should be used. In the analyses presented in Chapters 4 and 8, both left and right context were used. However, the psychological plausibility of using right context is, of course, open to question. Further empirical analyses could straightforwardly be carried out to assess the effects of removing this source of information. Regarding the question of the optimum window length, it was established in Chapter 5 that, for the 'standard' vector analyses carried out in Chapter 4, a relatively short window length of between 2 and 5 words in length produces the closest match to the categories contained within Roget's Thesaurus. Similar analyses might be conducted with the neural network introduced in Chapter 7 to determine whether this is also true when dealing with the more complex situation in which each target word can have more than a single representation. Once an optimum window length has been confirmed, a further form of lesioning might be introduced in which the window is reduced in length. The consequences of this, either during the training phase or during testing, for the representations developed for the target words would be of interest, perhaps indicating some of the consequences of reduced attentional or memory resources in using statistical information to learn the relationships between the meanings of words. It would also be worthy of consideration in the light of research which suggests that limited memory is important in the early stages of learning. Elman (1993), for example found that a recurrent neural network could learn a complex grammar only when the input was presented incrementally, starting with just a restricted amount of input information. The extent to which the network used in this thesis would permit relearning following damage is also of interest, because, as

Hinton and Shallice (1991) point out, both in their network and in some patients, relearning following a lesion takes place rapidly. As to the question of which type of context should be used, it is important to note that in the analyses carried out with the unsupervised neural network in Chapter 8, the context words were taken from the set of the most frequent 200 words in the Wall Street Journal corpus. This inevitably means that the target words were being represented largely in terms of closed class words, since this type of words dominates the most frequent 200 items. Whilst we have attempted to avoid an unprincipled choice of context words in this thesis, the use of a set of context words other than those taken from the most frequent words in the corpus is likely to be of interest, possibly providing some implications for the importance of closed class words in our understanding of word meanings.

A final issue of relevance to the network we have considered concerns the number of output units which are used for clustering the input data. As we have seen in Chapter 8, there is some evidence that there is an optimum number of these units. However, there is undoubtedly much further work to be carried out in confirming this. Furthermore, it would be of interest to explore the capabilities of such a network when the output units themselves could be added as required, rather than being part of a set whose numbers are predetermined from the start. In other words, rather than deciding at the outset that the network will contain, say, 30 output units, the decision as to the number of units needed would be left to the network to decide. 'Constructivist' approaches of this kind have attracted some interest within the general field of neural networks; Quartz and Sejnowski (1995), for example, support such approaches by noting that they appear to be important in the human brain.

9.2 Final Conclusions

In this thesis, we have examined various methods of allowing statistical structure present in English language data to determine a categorization of word meanings. Whilst we saw in Chapters 4 and 5 that 'standard' statistical methods can allow rich structures to be built up using this information alone, we also saw in Chapter 6 that such methods do suffer from various limitations. In particular, their provision for only a single, averaged, representation for each word was seen as inappropriate. In order

to investigate the ease with which this limitation might be overcome, an unsupervised neural network was introduced in Chapter 7 which could, in principle, allow statistical context to dictate the different senses into which words should be placed. In practice, the network performed well when dealing with Elman's (1988) corpus, and subsequently produced interesting results when applied to a large natural language corpus, as we saw in Chapter 8.

There is, as noted above, considerable scope for further exploration to be carried out in examining the usefulness of statistical structure in categorizing word meanings. However, we have nonetheless gone a considerable way towards satisfying the aims set out at the beginning of the thesis.

We noted initially that statistical information is a plausible candidate for informing a system which must learn the relationships between the meanings of different words. Whilst the process of acquiring word meanings has been of interest to psychologists for many years, it was suggested that the empirical work carried out within the field of Computational Linguistics might usefully be applied to the task of understanding how the process is carried out by human beings. We saw that, whilst extralinguistic factors are likely to be of importance, there are reasons to suppose that the intralinguistic information must also be a crucial influence in language acquisition. Our intention was to explore the *extent* to which such information might be useful, using minimal assumptions about the capabilities of the system carrying out the task. To this end, 'standard' analyses were carried out, in which each word is represented simply as a vector of conditional probabilities reflecting the statistical characteristics of the contexts in which the word occurs. Not even order information was recorded in these analyses. The results showed, nonetheless, that rich structures with an intuitively familiar appearance could be obtained through these methods. We subsequently made these intuitive assessments more objective and noted that there do appear to be optimal parameter settings when collecting statistical information about context. In particular, statistical information was seen not to be increasingly informative as the window length parameter was progressively increased.

These encouraging empirical analyses show that statistical structure is a potentially important source of information about word meanings, and confirm that we *can* begin with very local sorts of relationships when building up structures which reflect the similarities and differences between the meanings of words. They were then followed by the more desirable approach embodied in an unsupervised neural network. This was found to be capable of assigning words to clusters in an on-line fashion, with little prior information being supplied about the characteristics of the input domain. Even when faced with the complex problem of distributing words from a real corpus probabilistically over a number of output clusters, interesting structures were again observed.

Of course, the structures obtained with the various analyses carried out were not perfect, and contained numerous examples of word groupings which are not in accordance with our intuitions or with a benchmark such as Roget's Thesaurus. We may regard these analyses as an early step in investigating semantic categorization; to the extent that they produced 'incorrect' groupings, two possibilities for further enquiry emerge. Firstly, it may be that future refinement of the procedures used may enable more psychologically realistic structures to emerge. However, it is important to bear in mind that a priority in this thesis was to build in as few assumptions and as few *ad hoc* modifications as possible. It is to be hoped that any future work would also make this a priority. The second possibility is that we may find that there are ultimate limits on the informativeness of statistical structure in learning word meanings, and that any deficiencies in analyses such as those carried out here may be due to the fact that we have not included other, extralinguistic, sources of information. It seems highly likely that this will be the case, although it is nonetheless an empirical question to determine just how far a complicated system such as the human brain might proceed without the provision of extralinguistic information.

Whilst the work that has been carried out in this thesis does not, of course, *necessarily* indicate that knowledge of the statistical structure of language plays a major role in the acquisition of word meanings, it does provide substantial evidence that this could be the case. We noted in Chapter 2 that there are reasons for believing

that supervised learning could not account for the acquisition of a large part of our knowledge of word meanings, and saw that there is some experimental evidence to show that children may learn word meanings solely from hearing words used by those around them, and without supervision. In addition, we now have evidence that very simple unsupervised systems can go a considerable way towards producing a structured representation for word meanings without being given any information other than that which is present in the language data itself.

10. REFERENCES

Agarwal, R. (1995) Evaluation of Semantic Clusters. *Proceedings of the Association for Computational Linguistics*, **33**, 284-286.

Armstrong, S.L., Gleitman, L.R., and Gleitman, H. (1983) What some concepts might not be. *Cognition*, **13**, 263-308.

Baddeley, A. (1990) *Human Memory*. Lawrence Erlbaum. Hove, U.K.

Bard, E.G., Shillcock, R.C. and Altman, G.T.M. (1988) The recognition of words after their acoustic offsets in spontaneous speech: Effects of subsequent context. *Perception and Psychophysics*, **44**, 395-408.

Barlow, H.B. (1989) Unsupervised Learning. *Neural Computation*, **1**, 295-311.

Becker, C.A. (1980) Semantic context effects in visual word recognition: an analysis of semantic strategies. *Memory and Cognition*, **8**, 493-512.

Brady, J. (1993) An examination of concept lattices, types, and functions in *Roget's International Thesaurus*. *Behaviour Research Methods, Instruments, & Computers*, **25(2)**, 328-332.

Brown, P.F., Della Pietra, S.A., Della Pietra, V.J., and Mercer, R.L. (1991) Word-Sense Disambiguation Using Statistical Methods. *Proceedings of the Association for Computational Linguisitics*, **29**, 264-270.

Brown, P.F., Lai, J.C., and Mercer, R.L. (1991) Aligning Sentences in Parallel Corpora. *Proceedings of the Association for Computational Linguistics*, **29**, 169-176.

Brown, P.F., Della Pietra, V.J., deSouza, P.V., Lai, J.C. and Mercer, R.L. (1992) Class-based *n*-gram models of natural language. *Computational Linguistics*, **18(4)**, 467-479.

Bullinaria, J.A. and Huckle, C.C. (1996) Modelling Lexical Decision Using Corpus Derived Semantic Vectors in a Connectionist Network. Unpublished manuscript.

Charniak, E. (1993) *Statistical Language Learning*. MIT Press. Cambridge, Massachusetts.

Chater, N. and Conkey, P. (1992) Finding Linguistic Structure with Recurrent Neural Networks. In *Proceedings of the 14th Annual Conference of the Cognitive Science Society of America,* 402-407. Erlbaum. Hillsdale, New Jersey.

Chomsky, N. (1965) *Aspects of the theory of syntax*. MIT Press. Cambridge, Massachusetts.

Clark, E.V. (1979) *The ontogenesis of meaning*. Akademische Verlagsgesellschaft Athenaion. Wiesbaden.

Collins, A.M. and Quillian, M.R. (1969) Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, **8**, 240-247.

Elman, J.L. (1988) Finding Structure in Time. *CRL Technical Report 8801*. Center for Research in Language, University of California, San Diego, U.S.A.

Elman, J.L. (1993) Learning and Development in Neural Networks - The Importance of Starting Small. *Cognition*, 48(1), 71-99.

Finch, S. and Chater, N. (1991) A Hybrid Approach to the Automatic Learning of Linguistic Categories. *AISB Quarterly*, **78**, 16-24.

Finch, S.P. and Chater, N.J. (1992a) Bootstrapping syntactic categories. In *Proceedings of the 14th Annual Conference of the Cognitive Science Society of America,* 820-825. Erlbaum. Hillsdale, New Jersey.

Finch, S.P. and Chater, N.J. (1992b) Bootstrapping Syntactic Categories Using Statistical Methods. In *Proceedings of the First SHOE Workshop*, 230-235. Institute for Language Technology and AI. Tilburg University, The Netherlands.

Fodor, J.A. (1983) *The modularity of mind: An essay on faculty psychology.* Bradford. Cambridge, Massachusetts.

Forster, K.I. (1981) Frequency blocking and lexical access: One mental lexicon or two? *Journal of Verbal Learning and Verbal Behavior*, **20**, 190-203.

Fries, C.C. (1952) *The Structure of English.* Harcourt, Brace, and Company. New York.

Furnas, G.W., Landauer, T.K., Gomez, L.M. and Dumais, S.T. (1987) The Vocabulary Problem in Human-System Communication. *Communications of the ACM*, **30 (11)**, 964-971.

Gale, W.A. and Church, K.W. (1991) A Program for Aligning Sentences in Bilingual Corpora. *Proceedings of the Association for Computational Linguistics*, **29**, 177-184.

Gale, W., Church, K.W., and Yarowsky, D. (1992) Estimating Upper and Lower Bounds on the Performance of Word-Sense Disambiguation Programs. *Proceedings of ACL-92*, 249-256.

Gallant, S.I. (1991) A Practical Approach for Representing Context and for Performing Word Sense Disambiguation Using Neural Networks. *Neural Computation*, **3**, 293-309.

Garnham, A. (1990) *Psycholinguistics*. Routledge. London.

Gerrig, R.J. and Littman, M.L. (1990) Disambiguation by community membership. *Memory and Cognition*, **18(4),** 331-338.

Gleitman, L. (1994) The structural sources of verb meanings. In P. Bloom (Ed.), *Language Acquisition: Core Readings*. MIT Press, Cambridge, Massachusetts.

Grefenstette, G. (1993) Evaluation techniques for automatic semantic extraction: comparing syntactic and window-based approaches. In *Workshop on Acquisition of Lexical Knowledge from Text*. SIGLEX/ACL. Columbus, Ohio.

Grossberg, S. (1987) Competitive learning: from interactive activation to adaptive resonance. *Cognitive Science*, **11**, 23-63.

Haegeman, L. (1992) *Introduction to Government and Binding Theory*. Blackwell. Oxford, UK.

Harris, M. (1992) *Language Experience and Early Language Development*. Lawrence Erlbaum. Hove, UK.

Harris, Z.S. (1954) Distributional Structure. *Word*, **10**, 146-162.

Hertz, J., Krogh, A., and Palmer, R.G. (1991) *Introduction to the Theory of Neural Computation*. Addison-Wesley. Redwood City, California.

Hinton, G.E., Plaut, D.C., and Shallice, T. (1993) Simulating Brain Damage. *Scientific American*, **269(4)**, 76-82.

Hinton, G.E. and Shallice, T. (1991) Lesioning an Attractor Network: Investigations of Acquired Dyslexia. *Psychological Review*, **98(1)**, 74-95.

Hogaboam, T.W. and Perfetti, C.A. (1975) Lexical ambiguity and sentenc comprehension. *Journal of Verbal Learning and Verbal Behavior*, **14**, 265-274.

Hopfield, J.J. (1982) Neural Networks and Physical Systems with Emergent Collective Computational Abilities. In J.A. Anderson and E. Rosenfeld (Eds.) *Neurocomputing: Foundations of Research.* MIT Press. Cambridge.

Huckle, C.C. (1995) Grouping Words Using Statistical Context. In *Proceedings of the Seventh Conference of the European Chapter of the Association for Computational Linguistics*. Morgan Kaufmann. San Francisco, California.

Hughes, J. (1994) *Automatically Acquiring a Classification of Words*. Ph.D. thesis, University of Leeds, England, U.K.

Jones, G.V. (1985) Deep Dyslexia, Imageability, and Ease of Predication. *Brain and Language*, **24**, 1-19.

Jusczyk, P.W. (1993) How Word Recognition May Evolve from Infant Speech Perception Capacities. In G.T.M. Altmann and R. Shillcock (Eds.) *Cognitive Models of Speech Processing: The Second Sperlonga Meeting*. Lawrence Erlbaum. Hove, UK.

Lakoff, G. (1989) Cognitive Models and Prototype Theory. In U. Neisser (Ed) *Concepts and Conceptual Development.* Cambridge University Press. New York, U.S.A.

Lange, T.E. (1992) Lexical and pragmatic disambiguation and re-interpretation in connectionist networks. *International Journal of Man-Machine Studies*, **36**, 191-220.

Lapointe, F.-J. and Legendre, P. (1995) Comparative tests for dendrograms: A comparative evaluation. *Journal of Classification* (in press).

Lenneberg, E.H. (1967) *Biological Foundations of Language*. John Wiley. New York.

Lorr, M. (1983) *Cluster Analysis for Social Scientists*. Jossey-Bass. San Francisco.

Lund, K., Burgess, C., and Atchley, R.A. (1995) Semantic and Associative Priming in High-Dimensional Semantic Space. In *Proceedings of the 17th Annual Conference of the Cognitive Science Society of America,* 660-665. Erlbaum. Hillsdale, New Jersey.

McRae, K. (1992) Correlated properties in artifact and natural kind concepts. In *Proceedings of the 14th Annual Conference of the Cognitive Science Society of America*, 349-354, Erlbaum. Hillsdale, New Jersey.

McRae, K., de Sa, V.R. and Seidenberg, M.S. (1993) Modeling Property Intercorrelations in Conceptual Memory. In *Proceedings of the 15th Annual Conference of the Cognitive Science Society of America*, 729-734, Erlbaum. Hillsdale, New Jersey.

MacWhinney, B. and Snow, C. (1985) The child language data exchange system. *Journal of Child Language*, **12**, 271-95.

Mandelbrot, B.B. (1983) *The fractal geometry of nature*. W.H. Freeman. New York.

Marslen-Wilson, W.D. (1987) Functional parallelism in spoken word-recognition. *Cognition*, **25**, 71-102.

Marslen-Wilson, W. (1989) Access and Integration. In W. Marslen-Wilson (Ed) *Lexical Representation and Process*. MIT. Cambridge, Massachusetts.

Medin, D.L. and Smith, E.E. (1984) Concepts and Concept Formation. *Annual Review of Psychology*, **35**, 113-138.

Mervis, C. B. (1989) Child-basic object categories and early lexical development. In U. Neisser (Ed) *Concepts and Conceptual Development.* Cambridge University Press. New York, U.S.A.

Miller, G.A. (1963) *Language and Communication.* McGraw-Hill. New York, U.S.A.

Morton, J. (1970) A Functional Model for Memory. In D.A. Normal (Ed) *Models of Human Memory.* Academic Press. New York, U.S.A.

Moss, H.E. and Marslen-Wilson, W.D. (1993) Access to Word Meanings During Spoken Language Comprehension: Effects of Sentential Semantic Context. *Journal of Experimental Psychology*, **19 (6)**, 1254-1276.

Murre, J.M.J., Phaf, R.H. and Wolters, G. (1992) CALM: Categorizing and Learning Module. *Neural Networks*, **5**, 55-82.

Murtagh, F.D. (1993) Cluster Analysis Using Proximities. In I. Van Mechelen, J. Hampton, R.S. Michalski and P. Theuns (Eds.) *Categories and Concepts*. Academic Press. London.

Onifer, W. and Swinney, D.A. (1981) Accessing lexical ambiguity during sentence comprehension: Effects of frequency of meaning and contextual bias. *Memory and Cognition*, **9**, 225-236.

Pereira, F., Tishby, N. and Lee, L. (1993) Distributional Clustering of English Words. In *Proceedings of the Association for Computational Linguistics*, **31**, 183-190. ACL.

Pinker, S. (1994) *The Language Instinct.* William Morrow. New York, USA.

Plaut, D.C., McClelland, J.L., Seidenberg, M.S. and Patterson, K. (1996) Understanding Normal and Impaired Word Reading: Computational Principles in Quasi-Regular Domains. *Psychological Review*, **103(1)**, 56-115.

Plaut, D.C. and Shallice, T. (1993) Deep Dyslexia: A Case Study of Connectionist Neuropsychology. *Cognitive Neuropsychology*, **10(5)**, 377-500.

Plumbley, M.D. (1991) On Information Theory and Unsupervised Neural Networks. *Cambridge University Engineering Department Technical Report CUED/F-INFENG/TR.78.*

Quartz, S.R. and Sejnowski, T.J. (1995) The Neural Basis of Cognitive Development: A Constructivist Manifesto. Submitted paper.

Redington, M., Chater, N.J., and Finch, S.P. (1993) Distributional Information and the Acquisition of Linguistic Categories: A Statistical Approach. In *Proceedings of the 15th Annual Conference of the Cognitive Science Society of America,* 848-853. Erlbaum. Hillsdale, New Jersey.

Redington, M., Chater, N.J., and Finch, S.P. (1995) The Potential Contribution of Distributional Information to Early Syntactic Category Acquisition. Submitted paper.

Redlich, A.N. (1993) Redundancy Reduction as a Strategy for Unsupervised Learning. *Neural Computation*, **5**, 289-304.

Rice, M.L. (1990) Preschoolers' QUIL: Quick Incidental Learning of Words. In G. Conti-Ramsden and C.E. Snow (Eds.) *Children's Language Volume 7.* Lawrence Erlbaum. Hillsdale, New Jersey.

Ritter, H. and Kohonen, T. (1989) Self-Organizing Semantic Maps. *Biological Cybernetics*, **61**, 241-254.

Roitblat, H.L. and von Fersen, L. (1992) COMPARATIVE COGNITION: Representations and Processes in Learning and Memory. *Annual Review of Psychology*, **43**, 671-710.

Rosch, E. (1973) On the internal structure of perceptual and semantic categories. In T.E. Moore (Ed.) *Cognitive Development and the Acquisition of Language.* Academic Press. New York.

Rumelhart, D.E., Hinton, G.E., and Williams, R.J. (1986) Learning Internal Representations by Error Propagation. In D.E. Rumelhart, J.L. McClelland and the PDP Research Group (Eds) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations.* MIT Press. Cambridge, Massachusetts.

Rumelhart, D.E. and McClelland, J.L. On Learning the Past Tenses of English Verbs. In D.E. Rumelhart, J.L. McClelland and the PDP Research Group (Eds) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 2: Psychological and Biological Models.* MIT Press. Cambridge, Massachusetts.

Rumelhart, D.E. and Zipser, D. (1985) Feature Discovery by Competitive Learning. *Cognitive Science*, **9**, 75-112.

Scholtes, J.C. (1991) Using Extended Kohonen-Feature Maps in a Language Acquisition Model. In *Proceedings of the Second Australian Conference on Neural Networks*, 38-43.

Schütze, H. (1992) Word Sense Disambiguation With Sublexical Representations. In *Workshop Notes, Statistically-Based NLP Techniques*, 109-113. AAAI

Schütze, H. (1993a) Part-of-speech-induction from scratch. In *Proceedings of the Association for Computational Linguistics*, **31**, 251-258.

Schütze, H. (1993b) Word Space. In S. J. Hanson, J. D. Cowan, and C. L. Giles (Eds), *Advances in Neural Information Processing Systems* **5**, 895-902. Morgan Kaufmann. San Mateo, California.

Schütze, H. (1995) Distributional Part-of-Speech Tagging. In *Proceedings of the Seventh Conference of the European Chapter of the Association for Computational Linguistics*. Morgan Kaufmann. San Francisco, California.

Schütze, H. and Pedersen, J. (1993) A Vector Model for Syntagmatic and Paradigmatic Relatedness. In *Proceedings of the 9th Annual Conference of the Waterloo Centre for the New OED and Text Research*. Oxford, England.

Sedivy, J.C., Tanenhaus, M.K., Eberhard, K., Spivey-Knowlton, M., and Carlson, G.N. (1995) Using Intonationally-Marked Presuppositional Information in On-Line Language Processing: Evidence from Eye Movements to a Visual Model. In *Proceedings of the 17th Annual Conference of the Cognitive Science Society of America*, 375-380. Erlbaum. Mahwah, New Jersey.

Seidenberg, M.S. and McClelland, J.L. (1989) A Distributed, Developmental Model of Word Recognition and Naming. *Psychological Review*, **96(4)**, 523-568.

Shannon, C.E. and Weaver, W. (1963) *The Mathematical Theory of Communication*. University of Illinois Press. Urbana and Chicago.

Shillcock, R., Hicks, J., Cairns, P., Levy, J. and Chater, N. (1995) A statistical analysis of an idealised phonological transcription of the London-Lund corpus. Submitted to *Computer Speech and Language*.

Siegel, S. and Castellan, N.J. (1988) *Nonparametric Statistics for the Behavioural Sciences*. McGraw-Hill, Singapore.

Simpson, G.B. (1981) Meaning dominance and semantic context in the processing of lexical ambiguity. *Journal of Verbal Learning and Verbal Behavior*, **20**, 120-136.

Sinclair, J. (1991) *Corpus, Concordance, Collocation.* Oxford University Press. Oxford.

Skinner, B.F. (1957) *Verbal Behavior.* Appleton-Century-Crofts. New York.

Smadja, F.A. (1989) Lexical Co-occurrence: The Missing Link. *Literary and Linguistic Computing*, **4(3)**, 163-168.

Sonaiya, R. (1991) Vocabulary Acquisition as a Process of Continuous Lexical Disambiguation. *International Review of Applied Linguistics*, **29**, 273-284.

Spivey-Knowlton, M., Tanenhaus, M., Eberhard, K., and Sedivy, J. (1995) Eye Movements Accompanying Language and Action in a Visual Context: Evidence Against Modularity. In *Proceedings of the 17th Annual Conference of the Cognitive Science Society of America,* 25-30. Erlbaum. Mahwah, New Jersey.

Swinney, D.A. (1979) Lexical access during sentence comprehension: (Re)consideration of context effects. *Journal of Verbal Learning and Verbal Behavior*, **18**, 645-659.

Tabossi, P. and Zardon, F. (1993) Processing Ambiguous Words in Context. *Journal of Memory and Language*, **32**, 359-372.

Tanenhaus, M.K. and Lucas, M.M. (1987) Context effects in lexical processing. *Cognition*, **25**, 213-234.

Van Mechelen, I., Hampton, J., Michalski, R.S. and Theuns, P. (1993) *Categories and Concepts*. Academic Press. London.

Waltz, D.L. and Pollack, J.B. (1985) Massively Parallel Parsing: A Strongly Interactive Model of Natural Language Interpretation. *Cognitive Science*, **9**, 51-74.

Wetter, T. and Nüse, R. (1992) Use of natural language for knowledge acquisition: Strategies to cope with semantic and pragmatic variation. *IBM Journal of Research and Development*, **36(3)**, 435-468.

Wittgenstein, L. (1953) *Philosophical Investigations*. Blackwell, Oxford.

Wolff, J.G. (1976) Frequency, Conceptual Structure and Pattern Recognition. *British Journal of Psychology*, **67**, 377-390.

Wolff, J.G. (1988) Learning Syntax and Meanings Through Optimization and Distributional Analysis. In Y.Levy, I.M. Schlesinger and M.D.S. Braine (Eds.) *Categories and Processes in Language Acquisition*. LEA. Hillsdale, New Jersey.

Yarowsky, D. (1992) Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora. In *Proceedings of COLING-92*, Nantes.

Zavrel, J. and Veenstra, J. (1995) The Language Environment and Syntactic Word-Class Acquisition. In *Proceedings of the Groningen Assembly on Language Acquistion*. Groningen University.

Zipf, G.K. (1935) *The Psycho-biology of Language*. Houghton Mifflin, Boston.

Zipf, G.K. (1945) The Meaning-Frequency Relationship of Words. *The Journal of General Psychology*, **33**, 251-256.

Zwitserlood, P. (1989) The locus of the effects of sentential-semantic context in spoken-word processing. *Cognition*, **32**, 25-64.

APPENDIX A: Dendrograms and Tables of Nearest Neighbours

This appendix contains dendrograms for the 1000 target words considered in analyses 1-12 of Chapter 4.

Tables are also given here for the 10 nearest neighbours for each of the 50 randomly chosen target words considered in analyses 2-12 of Chapter 4.

The 50 randomly chosen target words are indicated in the dendrograms with capital letters.

Figure A.1 below shows the dendrogram resulting from analysis 1 in Chapter 4.

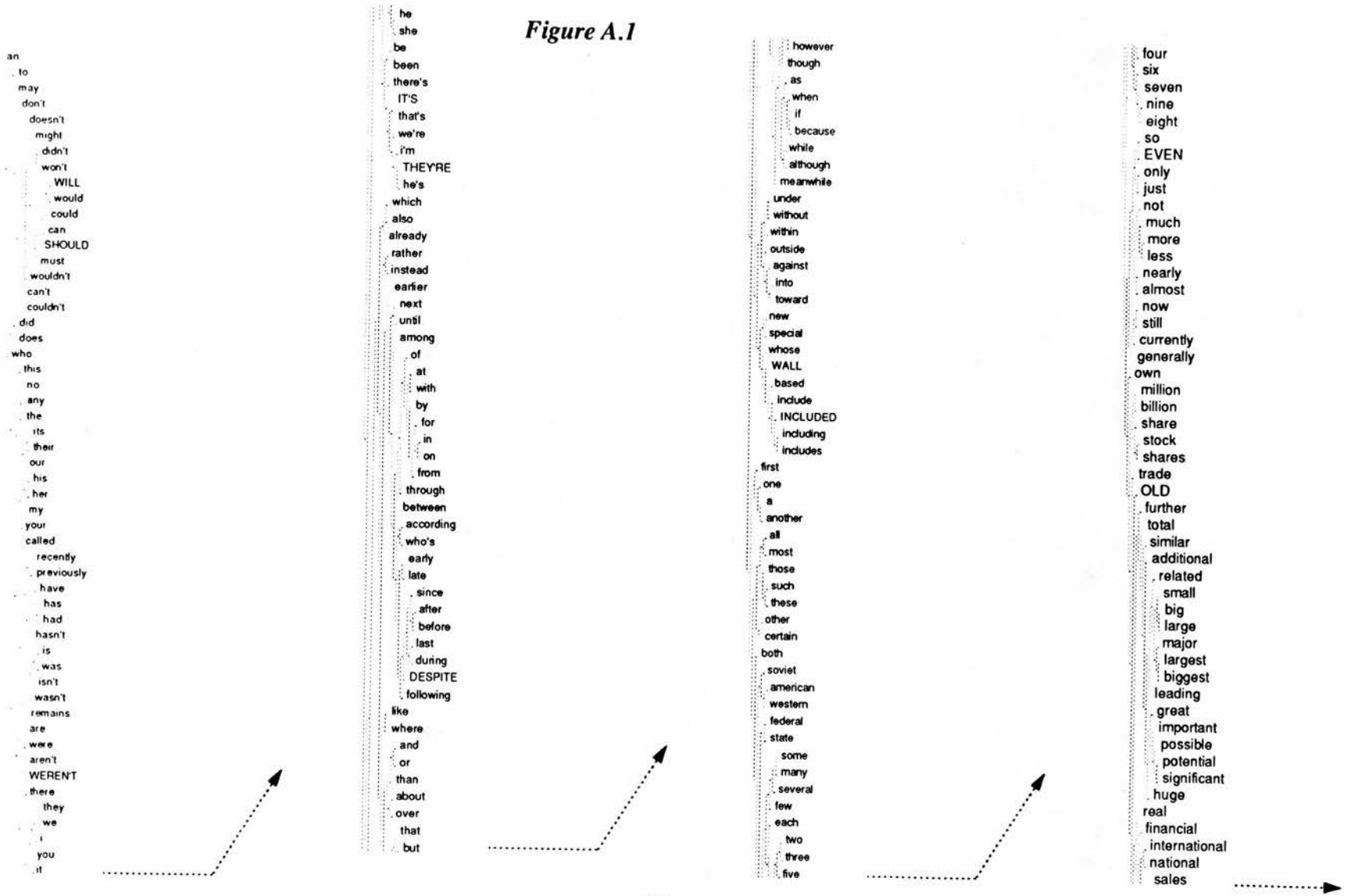

an
to
may
don't
doesn't
might
didn't
won't
WILL
would
could
can
SHOULD
must
wouldn't
can't
couldn't
did
does
who
this
no
any
the
its
their
our
his
her
my
your
called
recently
previously
have
has
had
hasn't
is
was
isn't
wasn't
remains
are
were
aren't
WEREN'T
there
they
we
i
you
it
he
she
be
been
there's
IT'S
that's
we're
i'm
THEY'RE
he's
which
also
already
rather
instead
earlier
next
until
among
of
at
with
by
for
in
on
from
through
between
according
who's
early
late
since
after
before
last
during
DESPITE
following
like
where
and
or
than
about
over
that
but
however
though
as
when
if
because
while
although
meanwhile
under
without
within
outside
against
into
toward
new
special
whose
WALL
based
include
INCLUDED
including
includes
first
one
a
another
all
most
those
such
these
other
certain
both
soviet
american
western
federal
state
some
many
several
few
each
two
three
five
four
six
seven
nine
eight
so
EVEN
only
just
not
much
more
less
nearly
almost
now
still
currently
generally
own
million
billion
share
stock
shares
trade
OLD
further
total
similar
additional
related
small
big
large
major
largest
biggest
leading
great
important
possible
potential
significant
huge
real
financial
international
national
sales

Figure A.1

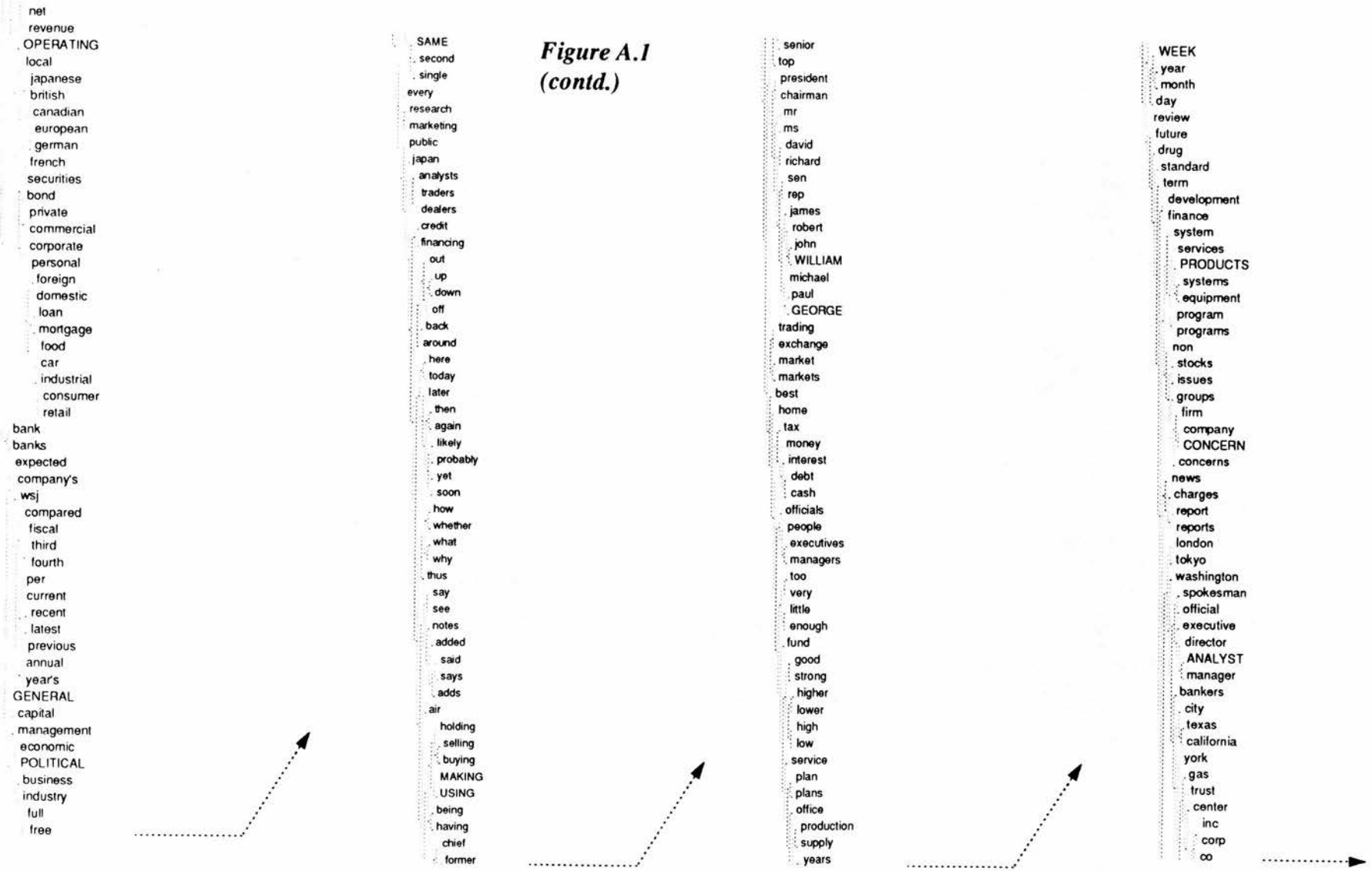


Figure A.1 (contd.)

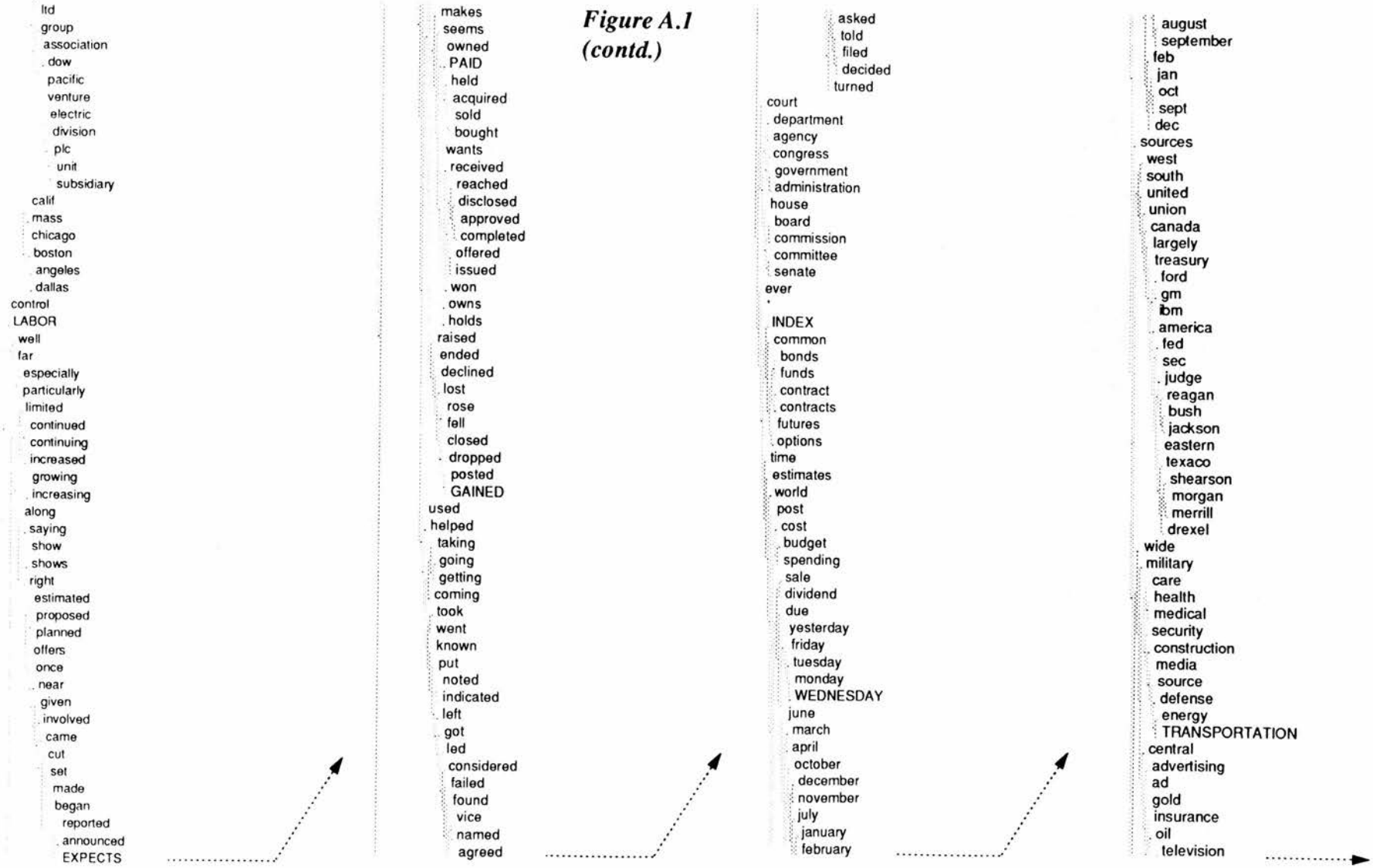


Figure A.1 (contd.)

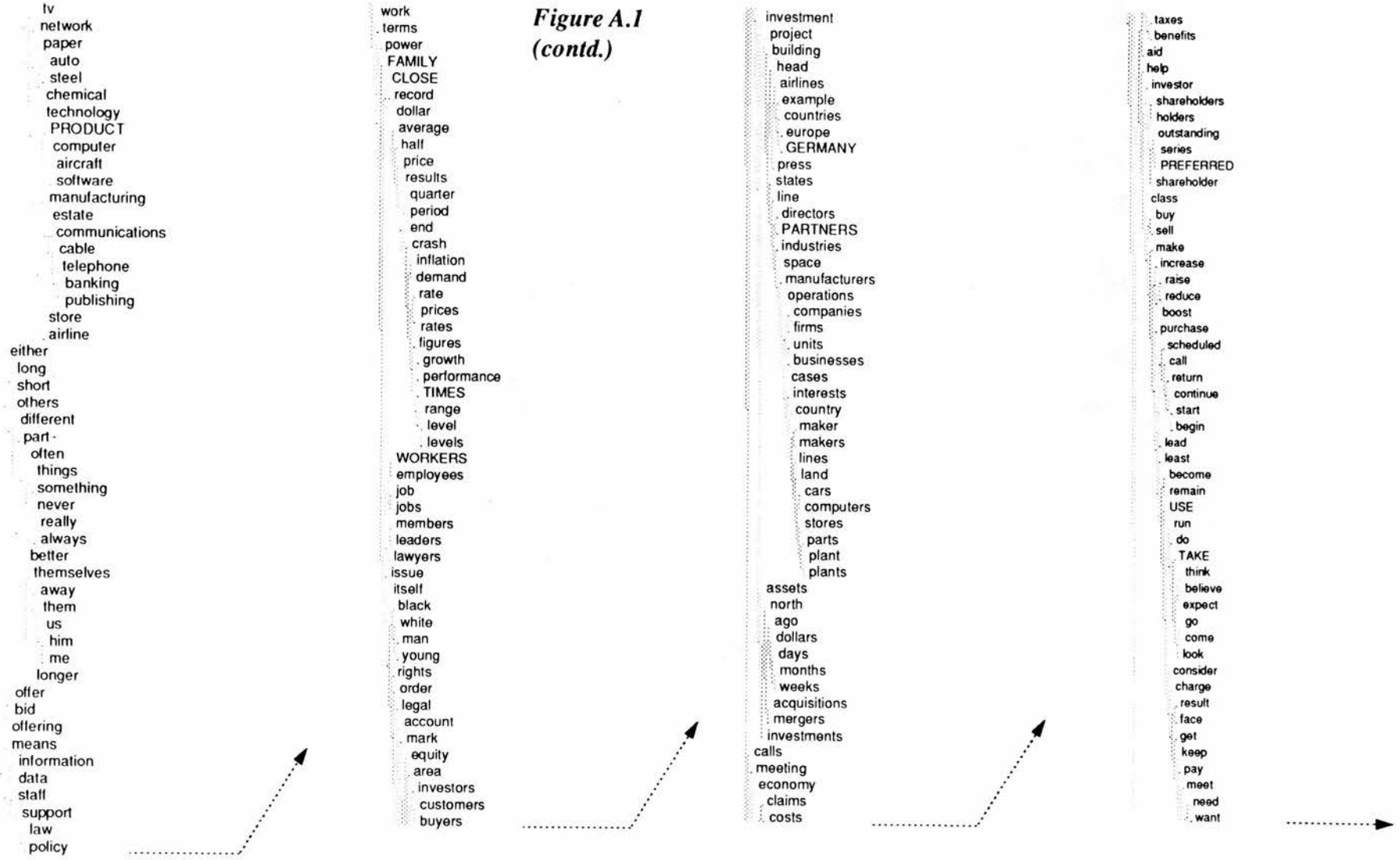


Figure A.1 (contd.)

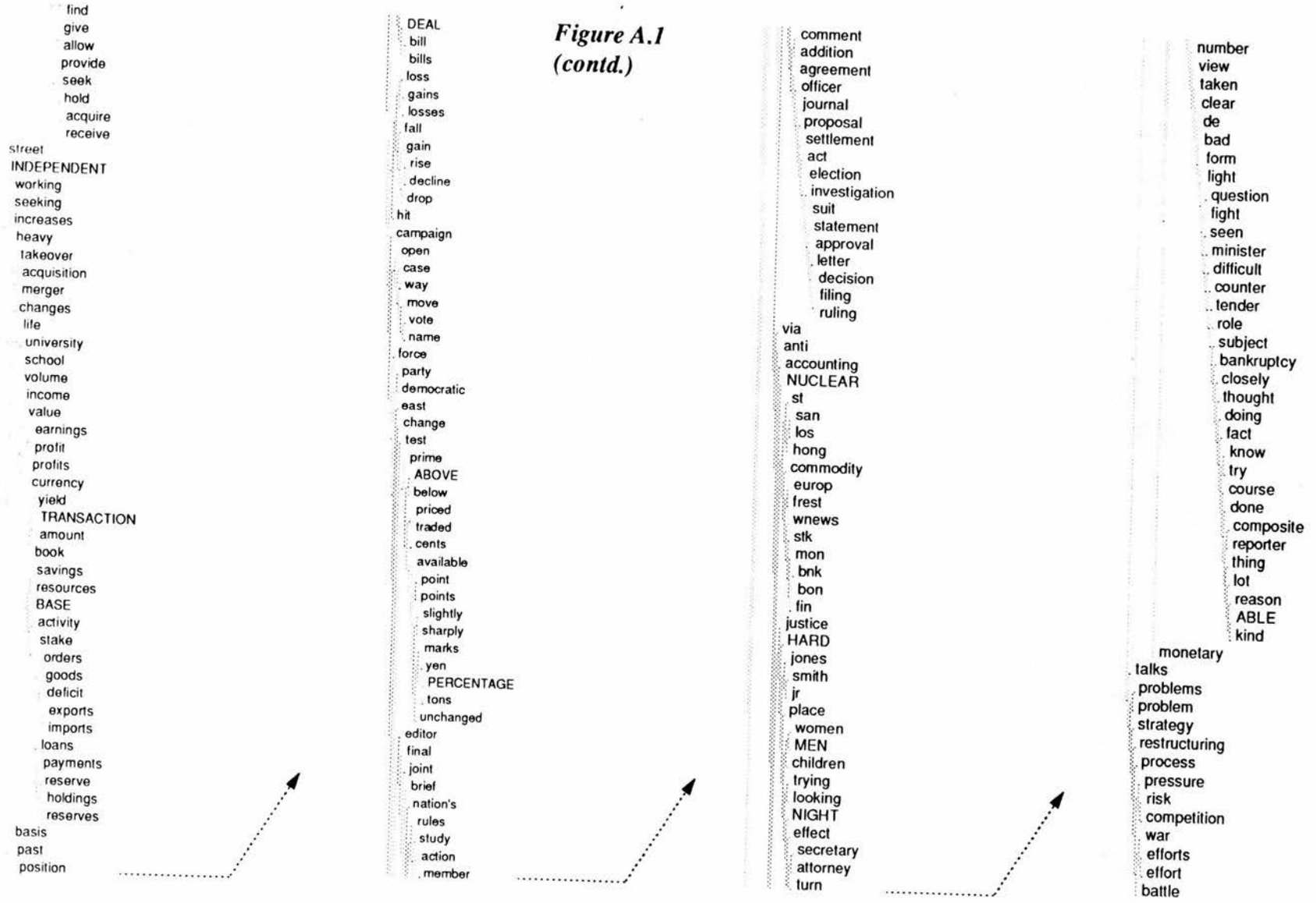


Figure A.1 (contd.)

Table A.2
(Euclidean Distance Metric, Window Length=1)

The following table contains the 50 target words and 10 nearest neighbours for analysis 2 in Chapter 4.

Target Word	10 Nearest Neighbours (Euclidean Distance)
able	according (0.283) trying (0.285) try (0.301) acquire (0.304) agreed (0.319) declined (0.323) going (0.336) continue (0.338) failed (0.340) sell (0.348)
above	over (0.097) below (0.101) current (0.105) biggest (0.107) best (0.108) under (0.108) project (0.109) during (0.109) despite (0.114) following (0.114)
analyst	estimated (0.210) additional (0.232) independent (0.238) official (0.247) investor (0.250) merrill (0.251) shearson (0.251) agreement (0.254) important (0.259) drexel (0.260)
base	public (0.091) price (0.092) non (0.092) tax (0.092) private (0.093) black (0.094) network (0.094) south (0.094) job (0.094) system (0.094)
close	keep (0.136) sell (0.138) failed (0.148) meet (0.153) raise (0.153) efforts (0.153) yield (0.160) help (0.162) plans (0.164) make (0.165)
concern	corp (0.077) ibm (0.086) co (0.087) inc (0.088) however (0.088) gm (0.090) it (0.091) mr (0.092) ford (0.092) officials (0.093)
deal	talks (0.114) agreement (0.126) test (0.131) hit (0.134) proposal (0.136) plan (0.137) name (0.139) leading (0.139) decision (0.141) move (0.141)
despite	on (0.051) after (0.059) under (0.059) in (0.060) into (0.061) french (0.065) before (0.066) around (0.067) against (0.068) through (0.068)
even	today (0.088) and (0.091) mr (0.094) they're (0.094) gm (0.096) ford (0.096) ms (0.096) he's (0.096) another (0.097) to (0.097)
expects	seems (0.123) wants (0.142) plans (0.156) make (0.178) decided (0.184) find (0.189) do (0.199) help (0.202) failed (0.203) get (0.206)
family	job (0.064) political (0.068) black (0.071) network (0.072) public (0.073) military (0.076) non (0.077) building (0.078) campaign (0.078) local (0.079)
gained	efforts (0.086) failed (0.089) rose (0.090) fell (0.094) continued (0.101) go (0.102) dropped (0.104) plans (0.105) decided (0.105) help (0.109)
general	non (0.079) political (0.081) management (0.082) ford (0.082) aircraft (0.083) co (0.083) black (0.084) credit (0.085) buying (0.085) legal (0.086)
george	mr (0.117) robert (0.119) ms (0.120) john (0.121) michael (0.121) james (0.121) david (0.124) richard (0.124) william (0.125) sen (0.128)
germany	german (0.111) point (0.302) cars (0.306) ford (0.306) businesses (0.306) operations (0.306) today (0.306) manufacturing (0.306) gm (0.307) stocks (0.307)
hard	me (0.094) continued (0.094) enough (0.098) dropped (0.102) aid (0.104) put (0.104) cut (0.104) run (0.106) see (0.106) related (0.111)
included	large (0.110) strong (0.117) good (0.117) different (0.118) private (0.120) little (0.123) major (0.125) small (0.125) very (0.128) issued (0.128)
independent	additional (0.099) investor (0.103) investment (0.116) equity (0.117) area (0.120) economic (0.122) office (0.125) aircraft (0.125) american (0.125)
index	prices (0.108) american (0.121) price (0.127) market (0.132) contracts (0.134) credit (0.136) and (0.137) issues (0.138) performance (0.138) funds (0.138)
it's	he's (0.057) that's (0.073) without (0.078) is (0.082) they're (0.088) strong (0.089) further (0.091) young (0.092) just (0.093) large (0.093)
labor	defense (0.069) military (0.080) by (0.082) meanwhile (0.084) political (0.084) for (0.084) that (0.085) while (0.085) system (0.085) public (0.086)
making	using (0.076) private (0.076) getting (0.077) building (0.077) non (0.077) growing (0.078) buying (0.081) today (0.082) made (0.083) black (0.083)
men	women (0.086) children (0.090) companies (0.090) groups (0.093) lawyers (0.097) workers (0.098) employees (0.098) people (0.099) aircraft (0.099) cars (0.101)
night	month (0.121) fall (0.127) friday (0.137) week (0.141) wednesday (0.143) until (0.151) tuesday (0.151) late (0.153) today (0.153) black (0.153)
nuclear	and (0.105) without (0.108) political (0.108) non (0.109) to (0.110) credit (0.110) co (0.110) black (0.112) meanwhile (0.113) william (0.114)
old	ago (0.120) five (0.154) last (0.163) earlier (0.182) seven (0.185) fiscal (0.186) this (0.190) three (0.193) four (0.193) eight (0.200)
operating	financial (0.174) and (0.182) mr (0.188) an (0.189) a (0.189) whose (0.189) another (0.190) investment (0.190) taking (0.190) ms (0.190)

Table A.2 (contd.)

paid	sold (0.074) made (0.085) itself (0.095) themselves (0.097) either (0.098) offered (0.098) further (0.099) another (0.099) limited (0.100) bought (0.105)
partners	today (0.066) and (0.069) ford (0.070) co (0.070) either (0.070) non (0.070) gm (0.071) programs (0.072) corp (0.074) management (0.074)
percentage	cost (0.184) total (0.185) level (0.187) basis (0.188) most (0.191) risk (0.193) top (0.194) unit (0.196) price (0.196) university (0.196)
political	legal (0.041) non (0.044) black (0.049) local (0.049) military (0.058) private (0.060) aircraft (0.062) management (0.064) building (0.065) young (0.067)
preferred	exchange (0.168) common (0.207) index (0.212) market (0.213) american (0.229) options (0.231) prices (0.232) price (0.241) markets (0.247) buying (0.251)
product	system (0.080) public (0.084) job (0.084) building (0.085) party (0.085) military (0.086) political (0.086) by (0.086) meanwhile (0.086) in (0.086)
products	businesses (0.066) programs (0.066) ford (0.068) gm (0.072) aircraft (0.072) groups (0.072) equipment (0.072) management (0.073) co (0.073) issues (0.074)
same	company's (0.117) nation's (0.134) past (0.144) latest (0.148) senate (0.221) company (0.221) sec (0.226) fed (0.228) dollar (0.228) during (0.235)
should	must (0.037) will (0.059) won't (0.062) may (0.069) might (0.070) would (0.072) could (0.075) can (0.083) couldn't (0.110) wouldn't (0.116)
take	get (0.059) provide (0.062) make (0.064) help (0.084) find (0.086) hold (0.091) go (0.091) give (0.092) pay (0.094) lead (0.098)
they're	we're (0.059) he's (0.062) mr (0.075) ms (0.077) are (0.082) and (0.083) that's (0.084) whose (0.086) were (0.087) it's (0.088)
times	problems (0.094) today (0.094) system (0.097) meanwhile (0.099) non (0.101) workers (0.101) off (0.101) that (0.102) south (0.102) programs (0.102)
transaction	during (0.079) economy (0.086) following (0.087) country (0.088) dollar (0.089) company (0.092) fed (0.095) senate (0.096) case (0.102) over (0.102)
transportation	defense (0.095) energy (0.110) construction (0.113) labor (0.115) aircraft (0.115) legal (0.115) political (0.116) steel (0.118) military (0.118) manufacturing (0.119)
use	purchase (0.089) start (0.096) support (0.103) review (0.106) test (0.112) act (0.117) cost (0.117) name (0.118) holders (0.119) units (0.121)
using	through (0.069) for (0.070) by (0.070) in (0.073) with (0.073) toward (0.073) british (0.074) after (0.074) public (0.075) only (0.075)
wall	journal (0.170) firms (0.472) in (0.481) by (0.481) for (0.481) of (0.481) toward (0.482) against (0.482) while (0.482) meanwhile (0.482)
wednesday	tuesday (0.041) monday (0.065) friday (0.076) feb (0.087) today (0.089) jan (0.091) sept (0.093) data (0.093) aircraft (0.096) programs (0.096)
week	month (0.067) fall (0.139) night (0.141) year (0.142) year's (0.142) later (0.192) friday (0.197) wednesday (0.197) until (0.198) time (0.199)
weren't	wasn't (0.213) aren't (0.214) were (0.215) are (0.215) and (0.222) mr (0.222) whose (0.223) ms (0.224) to (0.226) but (0.227)
will	won't (0.040) could (0.045) would (0.047) must (0.047) might (0.057) may (0.057) should (0.059) can (0.065) wouldn't (0.090) sold (0.113)
william	john (0.033) robert (0.034) david (0.036) michael (0.040) richard (0.041) james (0.042) ms (0.055) mr (0.058) sen (0.065) paul (0.068)
workers	employees (0.050) customers (0.063) ford (0.065) leaders (0.066) programs (0.067) today (0.068) operations (0.068) stocks (0.068) businesses (0.068) cars (0.069)

Figure A.2 below shows the dendrogram containing the 1000 target words considered in analysis 2 of Chapter 4.

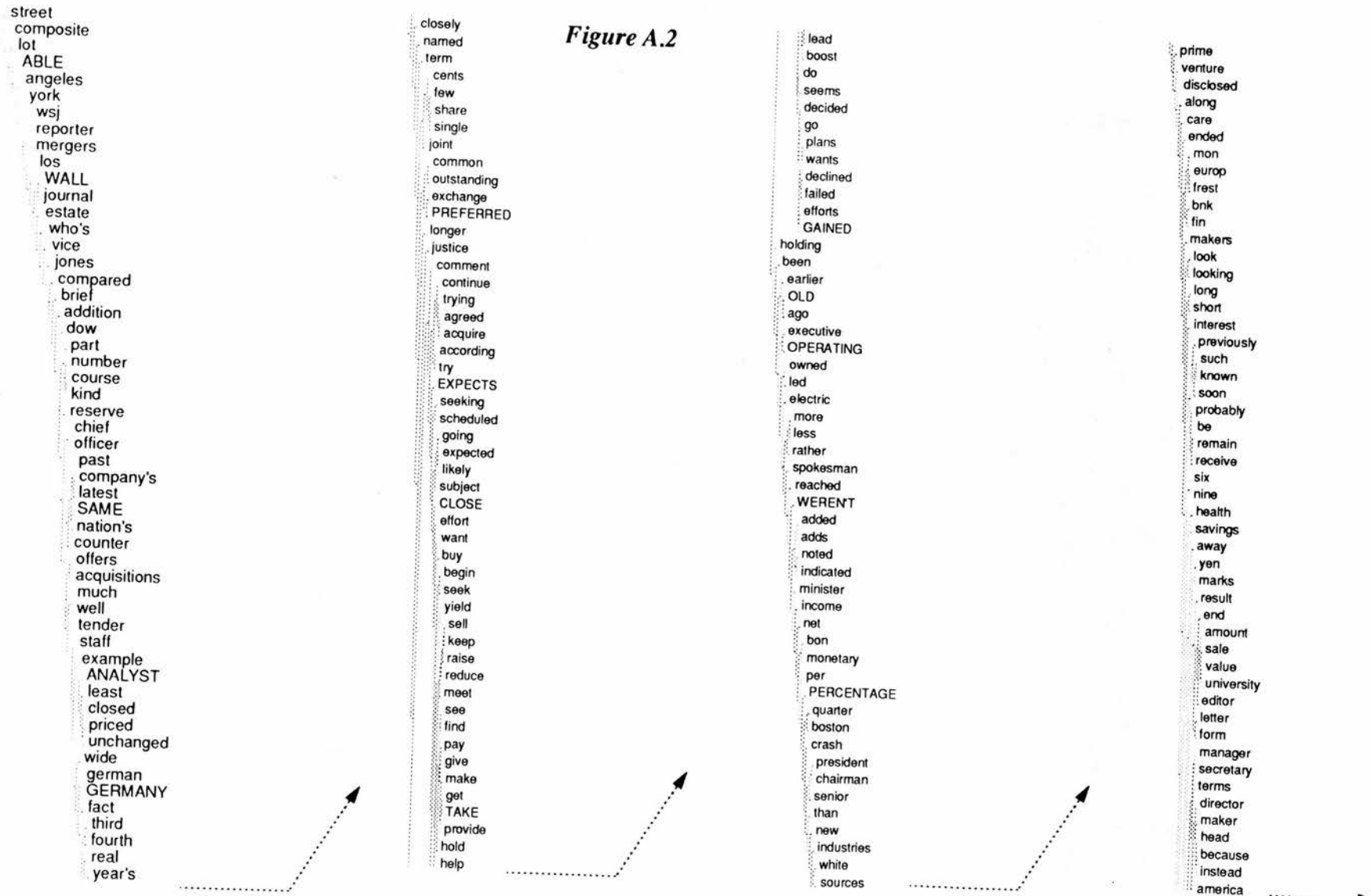

street
composite
lot
ABLE
angeles
york
wsj
reporter
mergers
los
WALL
journal
estate
who's
vice
jones
compared
brief
addition
dow
part
number
course
kind
reserve
chief
officer
past
company's
latest
SAME
nation's
counter
offers
acquisitions
much
well
tender
staff
example
ANALYST
least
closed
priced
unchanged
wide
german
GERMANY
fact
third
fourth
real
year's
closely
named
term
cents
few
share
single
joint
common
outstanding
exchange
PREFERRED
longer
justice
comment
continue
trying
agreed
acquire
according
try
EXPECTS
seeking
scheduled
going
expected
likely
subject
CLOSE
effort
want
buy
begin
seek
yield
sell
keep
raise
reduce
meet
see
find
pay
give
make
get
TAKE
provide
hold
help
lead
boost
do
seems
decided
go
plans
wants
declined
failed
efforts
GAINED
holding
been
earlier
OLD
ago
executive
OPERATING
owned
led
electric
more
less
rather
spokesman
reached
WEREN'T
added
adds
noted
indicated
minister
income
net
bon
monetary
per
PERCENTAGE
quarter
boston
crash
president
chairman
senior
than
new
industries
white
sources
prime
venture
disclosed
along
care
ended
mon
europ
frest
bnk
fin
makers
look
looking
long
short
interest
previously
such
known
soon
probably
be
remain
receive
six
nine
health
savings
away
yen
marks
result
end
amount
sale
value
university
editor
letter
form
manager
secretary
terms
director
maker
head
because
instead
america

Figure A.2

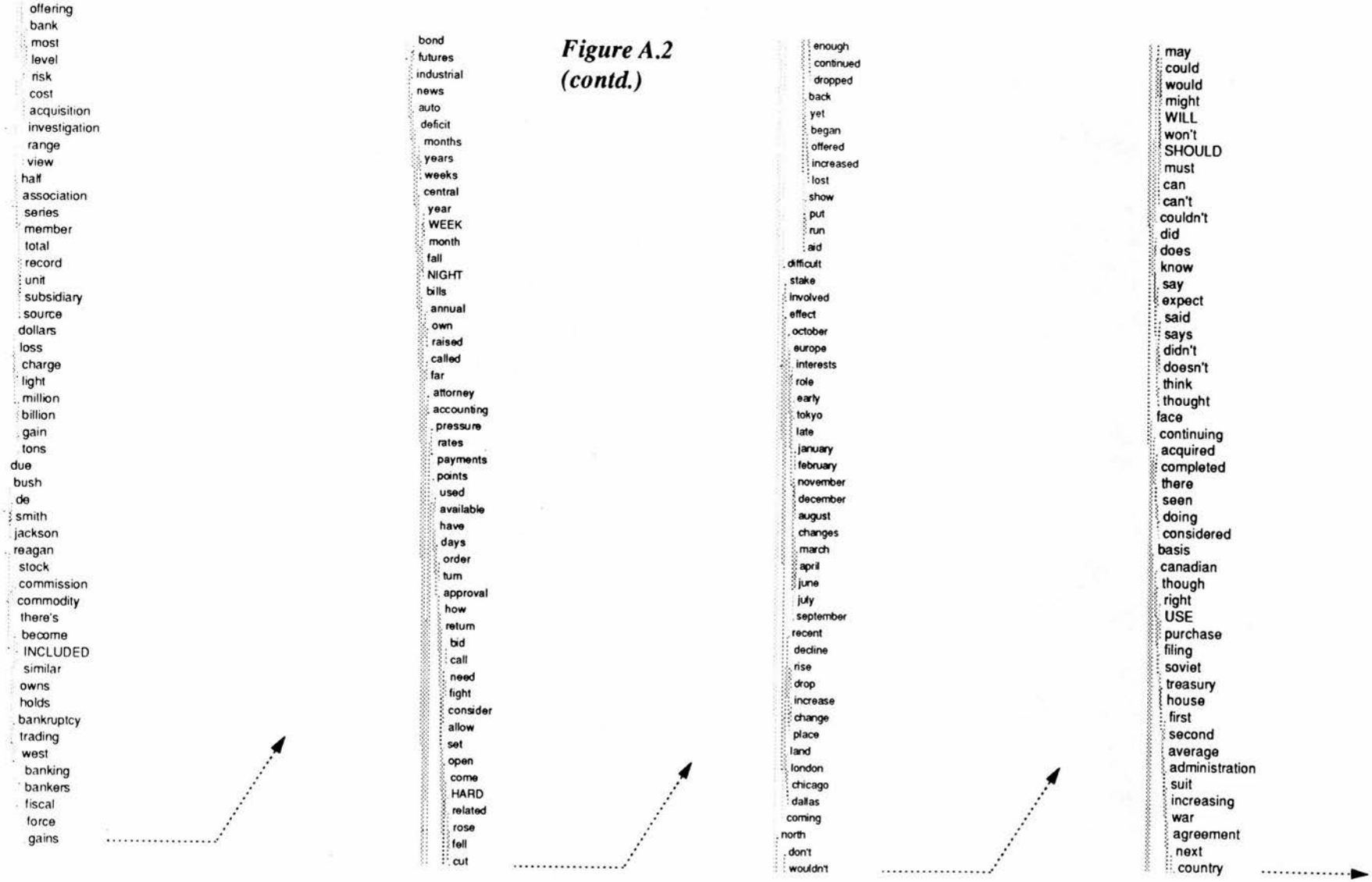


Figure A.2 (contd.)

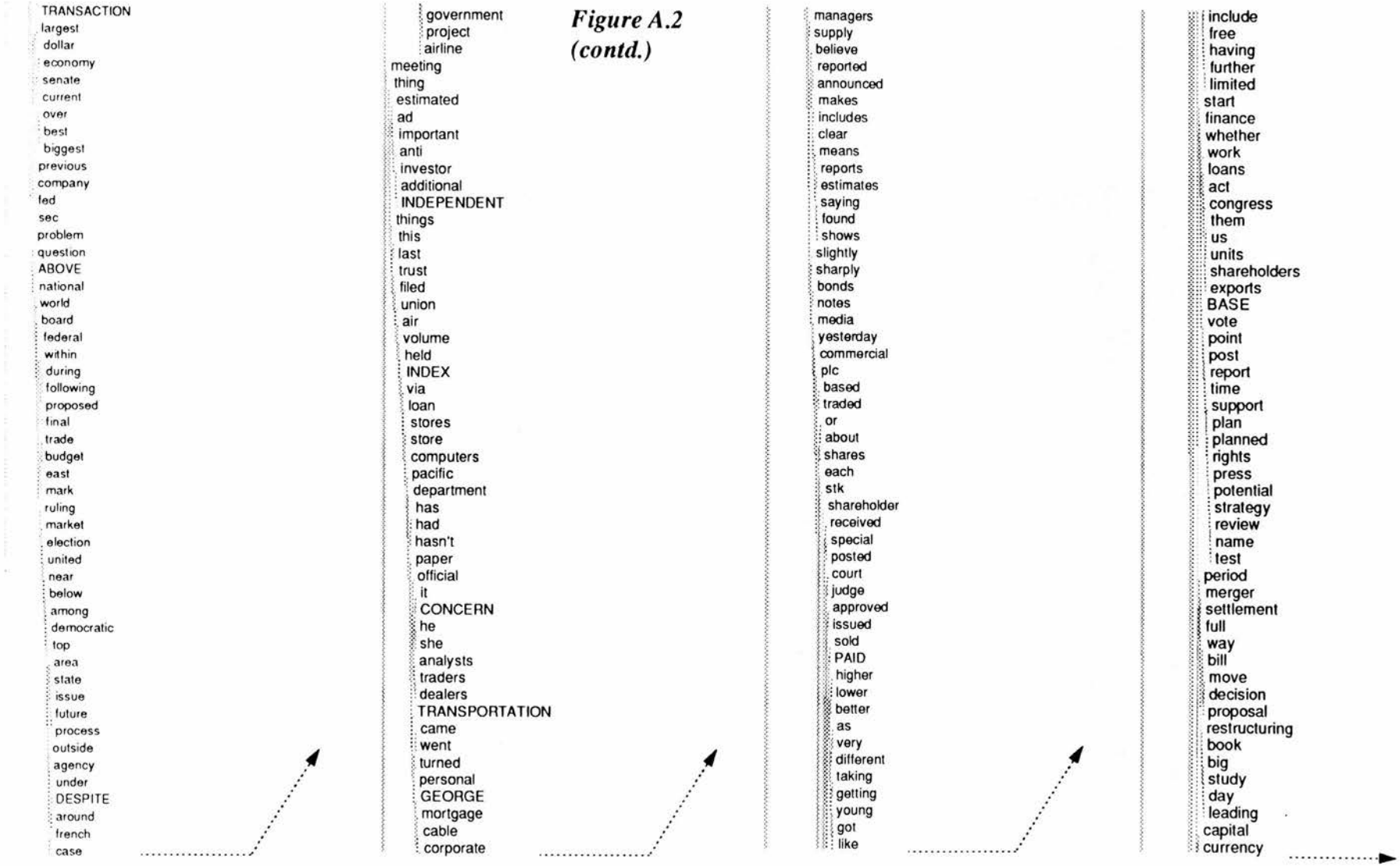


Figure A.2 (contd.)

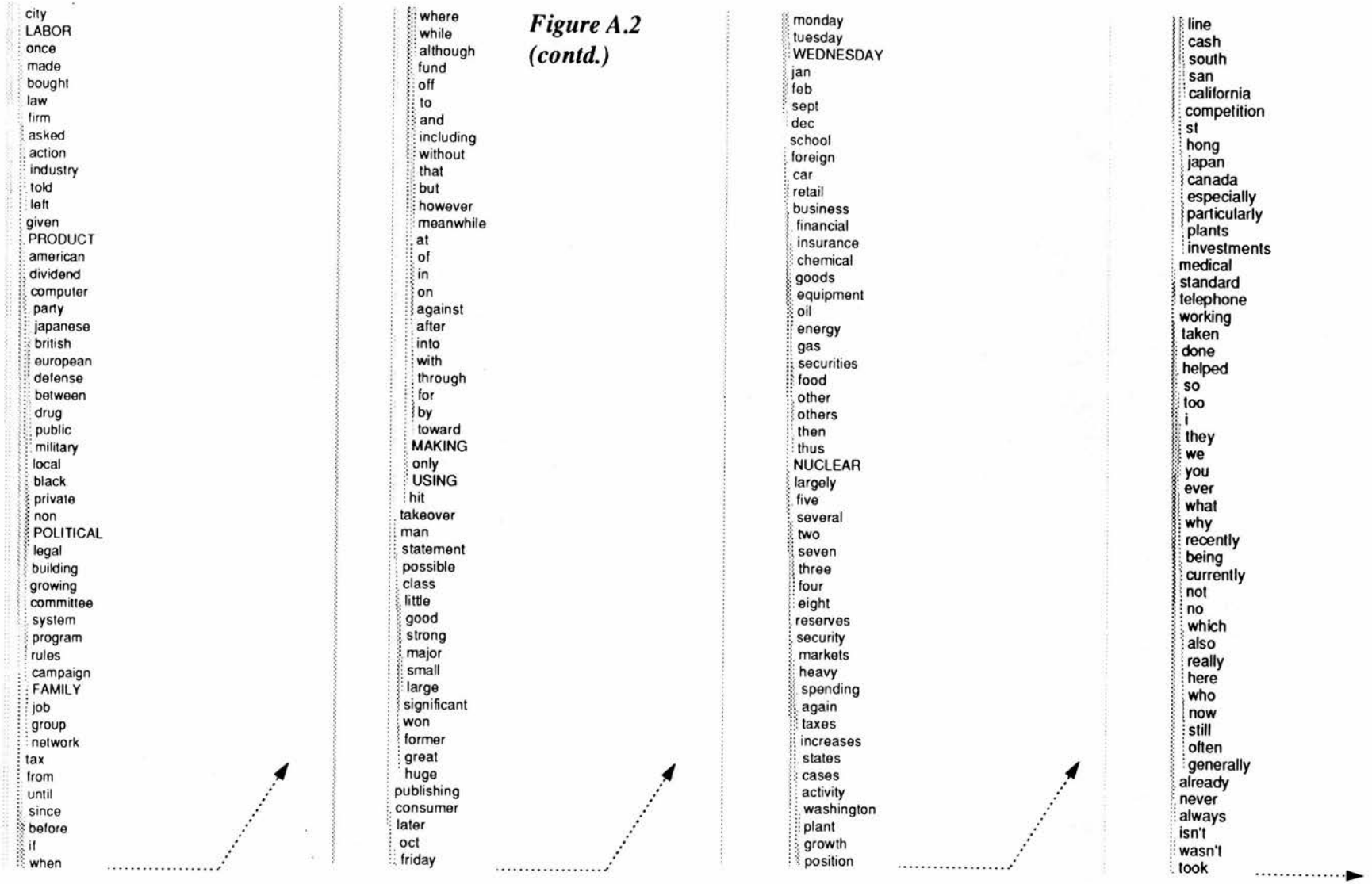


Figure A.2 (contd.)

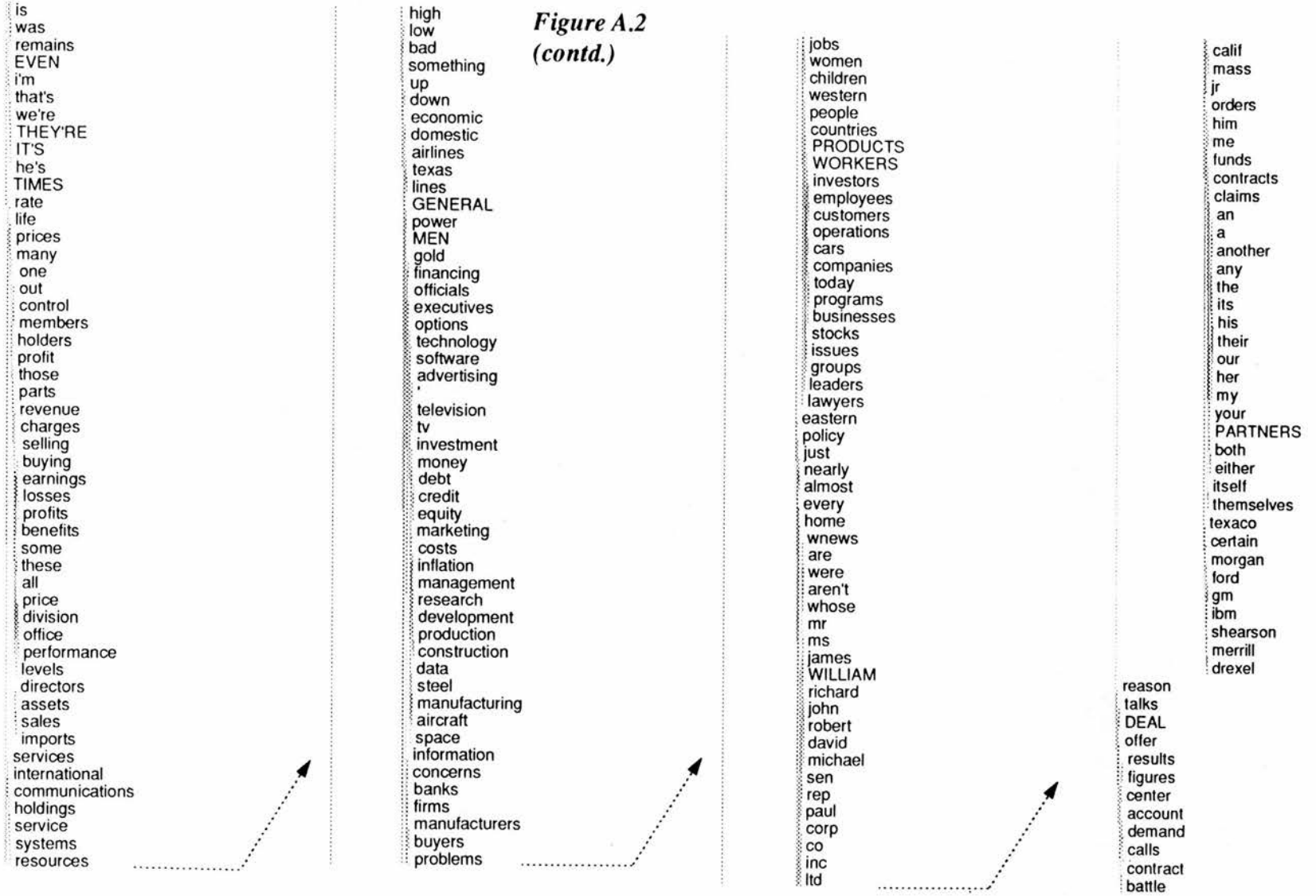


Figure A.2 (contd.)

Table A.3
(Spearman Distance Metric, Window Length=2)

The table below contains the 50 target words and 10 nearest neighbours for analysis 3 in Chapter 4.

Target Word	10 Nearest Neighbours (Spearman Correlation Coefficient)
able	trying (0.600) try (0.579) want (0.571) decided (0.566) likely (0.562) going (0.561) difficult (0.543) them (0.540) they (0.540) can (0.540)
above	below (0.671) down (0.608) up (0.603) higher (0.575) than (0.569) lower (0.562) about (0.557) percentage (0.550) from (0.550) year (0.548)
analyst	reporter (0.517) manager (0.507) director (0.496) said (0.485) counter (0.480) analysts (0.479) inc (0.474) and (0.474) gained (0.473) ltd (0.471)
base	and (0.509) costs (0.504) growth (0.501) system (0.495) network (0.494) lines (0.493) space (0.493) cost (0.491) demand (0.489) reserves (0.487)
close	down (0.538) monday (0.527) point (0.514) but (0.513) dollar (0.513) up (0.510) traders (0.507) dealers (0.507) bid (0.502) marks (0.501)
concern	company (0.692) group (0.621) firm (0.612) concerns (0.609) unit (0.604) subsidiary (0.593) maker (0.585) division (0.583) inc (0.577) operations (0.574)
deal	thing (0.568) do (0.542) out (0.535) he (0.531) so (0.530) work (0.529) know (0.529) something (0.528) it (0.528) even (0.526)
despite	in (0.603) after (0.589) while (0.586) on (0.585) from (0.572) during (0.571) as (0.551) because (0.549) following (0.539) but (0.533)
even	so (0.723) but (0.721) still (0.676) they (0.673) if (0.668) though (0.664) much (0.657) because (0.653) some (0.652) we (0.650)
expects	expect (0.556) expected (0.545) will (0.539) posted (0.524) reported (0.494) plans (0.491) declined (0.485) would (0.479) for (0.472) and (0.471)
family	life (0.551) children (0.536) care (0.520) men (0.514) working (0.511) reporter (0.510) home (0.507) women (0.502) employees (0.499) school (0.497)
gained	fell (0.657) rose (0.630) closed (0.628) dropped (0.623) posted (0.608) declined (0.593) unchanged (0.588) lost (0.572) traded (0.570) cents (0.560)
george	reporter (0.553) paul (0.535) mr (0.531) president (0.521) says (0.520) john (0.520) jr (0.519) robert (0.516) james (0.515)
germany	europe (0.575) japan (0.547) minister (0.546) economy (0.533) west (0.533) exports (0.531) reporter (0.525) leaders (0.517) east (0.516) countries (0.513)
general	new (0.541) international (0.535) national (0.522) group (0.521) an (0.515) said (0.513) american (0.513) communications (0.509) and (0.504) electric (0.503)
hard	people (0.599) them (0.591) difficult (0.589) so (0.588) not (0.587) we (0.584) too (0.582) they (0.575) what (0.574)
included	reported (0.521) include (0.516) from (0.511) includes (0.505) including (0.503) by (0.498) posted (0.491) for (0.491) increased (0.489) while (0.488)
independent	european (0.488) mergers (0.485) service (0.485) publishing (0.484) state (0.483) turned (0.475) agency (0.475) california (0.474) anti (0.473) owned (0.473)
index	futures (0.616) prices (0.602) stock (0.588) traders (0.582) stocks (0.577) market (0.571) average (0.570) volume (0.567) trading (0.551)
it's	that's (0.677) so (0.663) there (0.662) is (0.653) i (0.645) not (0.642) he's (0.640) but (0.635) i'm (0.634) they're (0.632)
labor	trade (0.502) air (0.499) journal (0.474) fin (0.472) justice (0.464) state (0.457) economic (0.457) health (0.455) union (0.454) industry (0.451)
making	made (0.622) make (0.604) using (0.558) because (0.556) but (0.554) as (0.553) taking (0.551) so (0.543) into (0.542) for (0.536)
men	people (0.670) women (0.661) children (0.610) them (0.595) me (0.581) him (0.569) things (0.568) man (0.566) something (0.565) thing (0.564)
night	mr (0.547) week (0.539) vote (0.534) he (0.532) reporter (0.521) thing (0.520) election (0.519) day (0.516) weeks (0.513) court (0.513)
nuclear	plant (0.538) aircraft (0.530) defense (0.529) military (0.526) power (0.519) space (0.517) plants (0.516) control (0.507) east (0.502) mergers (0.501)
old	says (0.545) who (0.533) years (0.525) reporter (0.525) i (0.506) left (0.503) like (0.502) man (0.501) head (0.498) young (0.492)
operating	profit (0.603) earnings (0.603) sales (0.601) net (0.591) revenue (0.553) quarter (0.549) fourth (0.539) income (0.536) year (0.534) fiscal (0.533)

Table A.3 (contd.)

paid	pay (0.609) sold (0.598) bought (0.578) offered (0.560) held (0.555) traded (0.554) and (0.536) received (0.532) lost (0.531) available (0.528)
partners	group (0.558) directors (0.545) firm (0.541) executives (0.524) companies (0.517) company (0.517) interests (0.506) stores (0.504) venture (0.504) reporter (0.503)
percentage	average (0.605) tons (0.595) points (0.581) per (0.578) volume (0.572) rate (0.568) slightly (0.566) yield (0.563) compared (0.561) below (0.560)
political	economic (0.597) democratic (0.580) legal (0.547) policy (0.539) party (0.534) military (0.527) way (0.526) leaders (0.517) election (0.515) campaign (0.511)
preferred	common (0.587) holders (0.576) shares (0.567) outstanding (0.565) debt (0.552) transaction (0.540) stock (0.539) share (0.532) shareholders (0.530) million (0.530)
product	products (0.578) cars (0.525) businesses (0.522) technology (0.520) software (0.518) store (0.514) strategy (0.511) computers (0.511) car (0.508) information (0.507)
products	equipment (0.627) systems (0.618) businesses (0.606) services (0.594) parts (0.590) product (0.578) technology (0.577) computers (0.576) computer (0.564) companies (0.560)
same	this (0.551) one (0.543) that (0.542) only (0.536) all (0.523) some (0.520) each (0.518) even (0.517) different (0.515) now (0.514)
should	would (0.780) could (0.756) will (0.746) can (0.715) must (0.708) might (0.698) may (0.654) won't (0.643) not (0.625) don't (0.592)
take	get (0.649) make (0.638) give (0.620) go (0.612) put (0.598) do (0.595) come (0.584) taking (0.576) find (0.575)
they're	it's (0.632) we're (0.616) i'm (0.610) he's (0.608) not (0.596) so (0.591) people (0.590) they (0.590) that's (0.577) now (0.577)
times	time (0.527) years (0.516) year (0.515) book (0.509) down (0.508) around (0.505) months (0.499) one (0.499) percentage (0.498) about (0.497)
transaction	offer (0.606) acquisition (0.583) sale (0.580) merger (0.556) bid (0.553) agreement (0.545) purchase (0.542) preferred (0.540) settlement (0.529) debt (0.526)
transportation	services (0.505) energy (0.499) food (0.497) insurance (0.495) health (0.493) electric (0.488) defense (0.482) development (0.479) telephone (0.476) equipment (0.476)
use	using (0.578) get (0.559) do (0.553) find (0.550) have (0.550) used (0.549) sell (0.546) provide (0.539) make (0.536) buy (0.527)
using	use (0.578) used (0.563) making (0.558) and (0.554) into (0.544) buying (0.538) other (0.534) such (0.527) selling (0.526) taking (0.519)
wall	street (0.752) its (0.560) the (0.535) said (0.510) and (0.510) posted (0.508) earlier (0.500) new (0.499) last (0.497) reported (0.493)
wednesday	friday (0.671) tuesday (0.647) monday (0.645) yesterday (0.629) week (0.603) last (0.582) late (0.580) april (0.568) june (0.560) march (0.558)
week	month (0.695) last (0.665) monday (0.649) year (0.644) friday (0.639) yesterday (0.623) tuesday (0.620) weeks (0.609) wednesday (0.603) april (0.601)
weren't	were (0.604) are (0.588) aren't (0.573) been (0.524) wasn't (0.501) had (0.494) be (0.485) terms (0.472) have (0.471) isn't (0.470)
will	would (0.866) could (0.804) should (0.746) may (0.736) won't (0.716) can (0.708) might (0.689) must (0.679) to (0.635) didn't (0.633)
william	robert (0.578) john (0.571) james (0.554) richard (0.546) david (0.544) paul (0.532) jr (0.521) michael (0.518) chairman (0.505) george (0.497
workers	employees (0.640) jobs (0.576) work (0.575) women (0.563) people (0.560) members (0.550) benefits (0.538) managers (0.534) states (0.531) companies (0.531)

Figure A.3 below shows the dendrogram containing the 1000 target words considered in analysis 3 of Chapter 4.

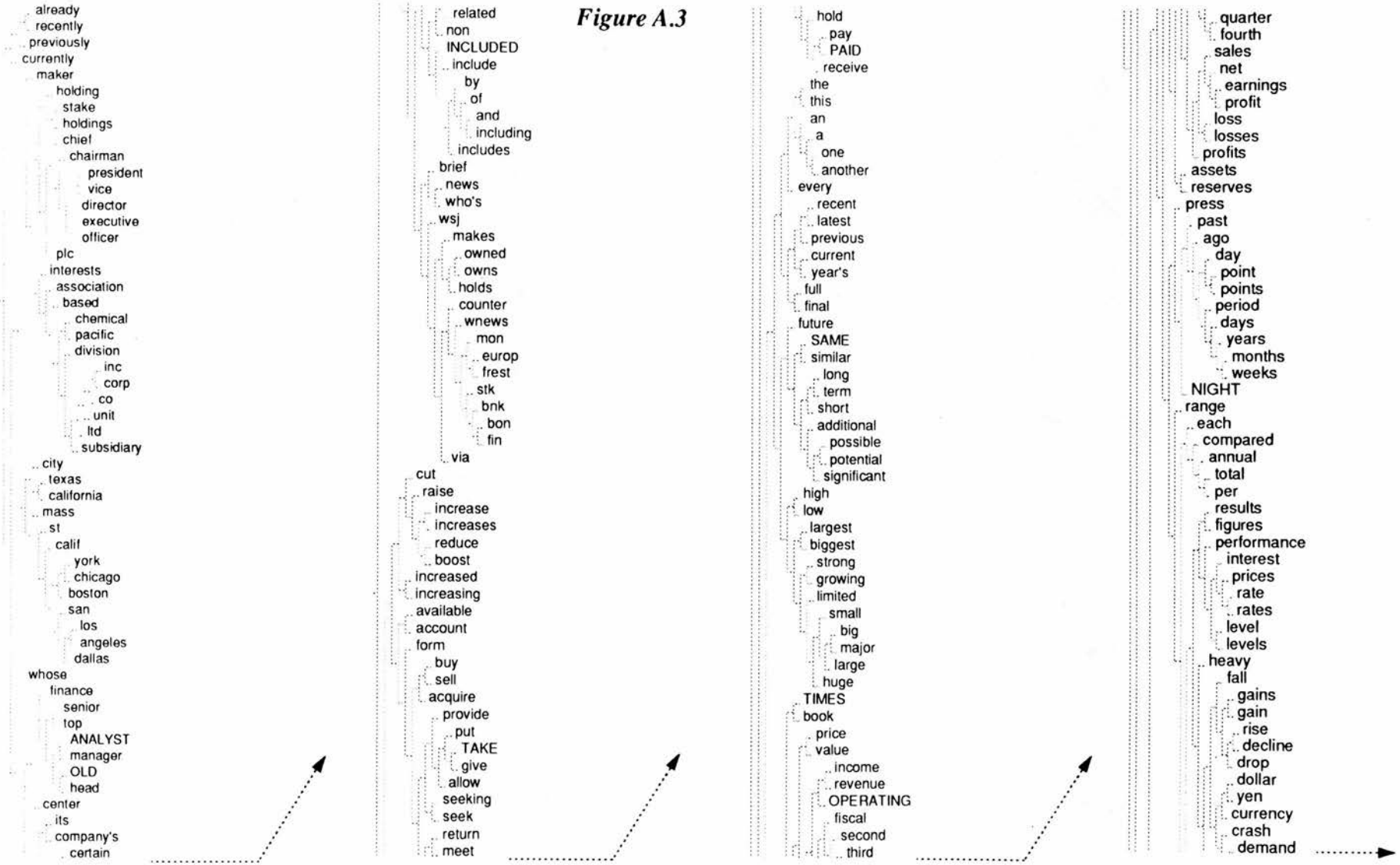

already
recently
previously
currently
maker
holding
stake
holdings
chief
chairman
president
vice
director
executive
officer
plc
interests
association
based
chemical
pacific
division
inc
corp
co
unit
ltd
subsidiary
city
texas
california
mass
st
calif
york
chicago
boston
san
los
angeles
dallas
whose
finance
senior
top
ANALYST
manager
OLD
head
center
its
company's
certain
related
non
INCLUDED
include
by
of
and
including
includes
brief
news
who's
wsj
makes
owned
owns
holds
counter
wnews
mon
europ
frest
stk
bnk
bon
fin
via
cut
raise
increase
increases
reduce
boost
increased
increasing
available
account
form
buy
sell
acquire
provide
put
TAKE
give
allow
seeking
seek
return
meet
hold
pay
PAID
receive
the
this
an
a
one
another
every
recent
latest
previous
current
year's
full
final
future
SAME
similar
long
term
short
additional
possible
potential
significant
high
low
largest
biggest
strong
growing
limited
small
big
major
large
huge
TIMES
book
price
value
income
revenue
OPERATING
fiscal
second
third
quarter
fourth
sales
net
earnings
profit
loss
losses
profits
assets
reserves
press
past
ago
day
point
points
period
days
years
months
weeks
NIGHT
range
each
compared
annual
total
per
results
figures
performance
interest
prices
rate
rates
level
levels
heavy
fall
gains
gain
rise
decline
drop
dollar
yen
currency
crash
demand

Figure A.3

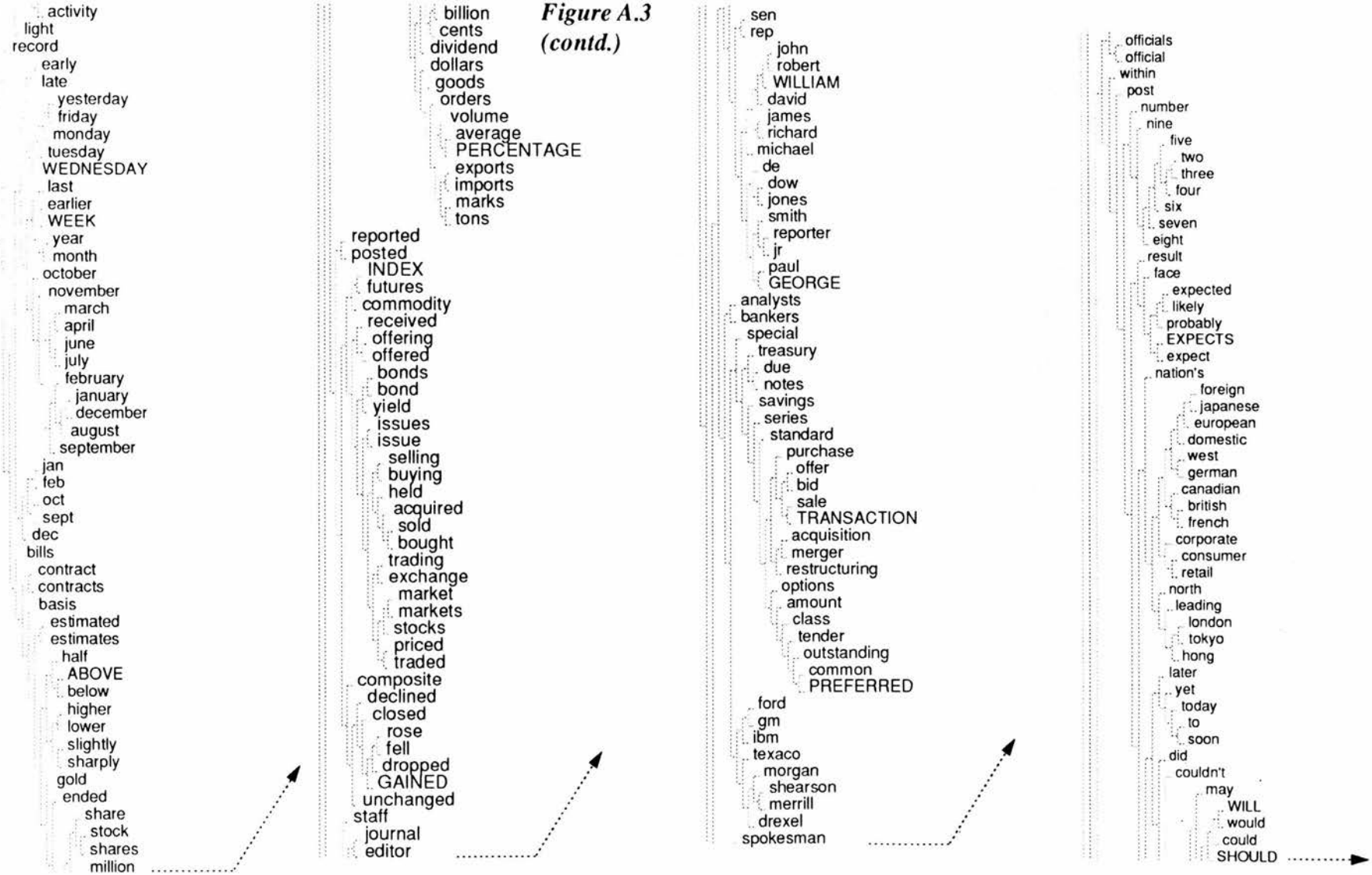


Figure A.3 (contd.)

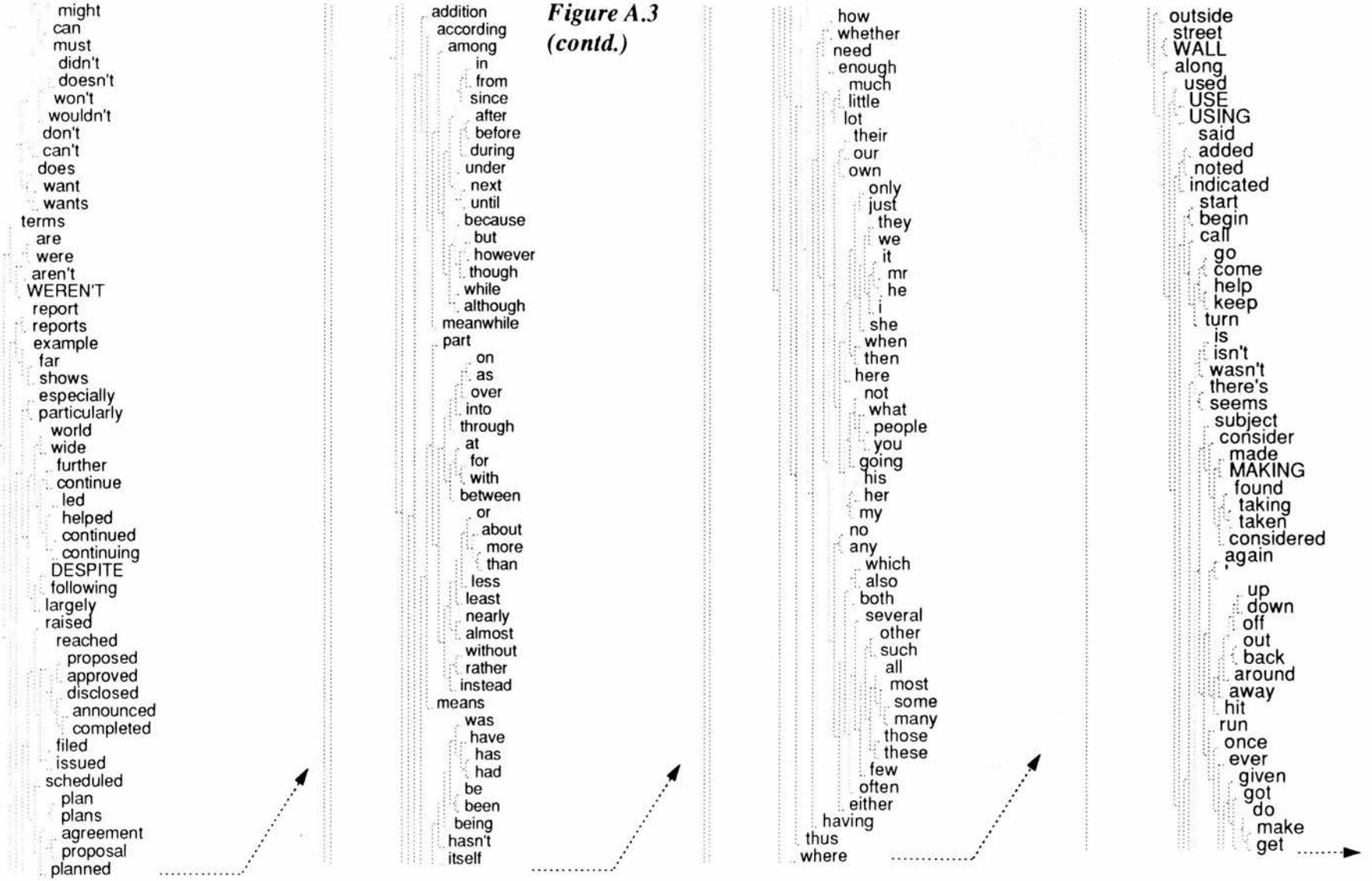


Figure A.3 (contd.)

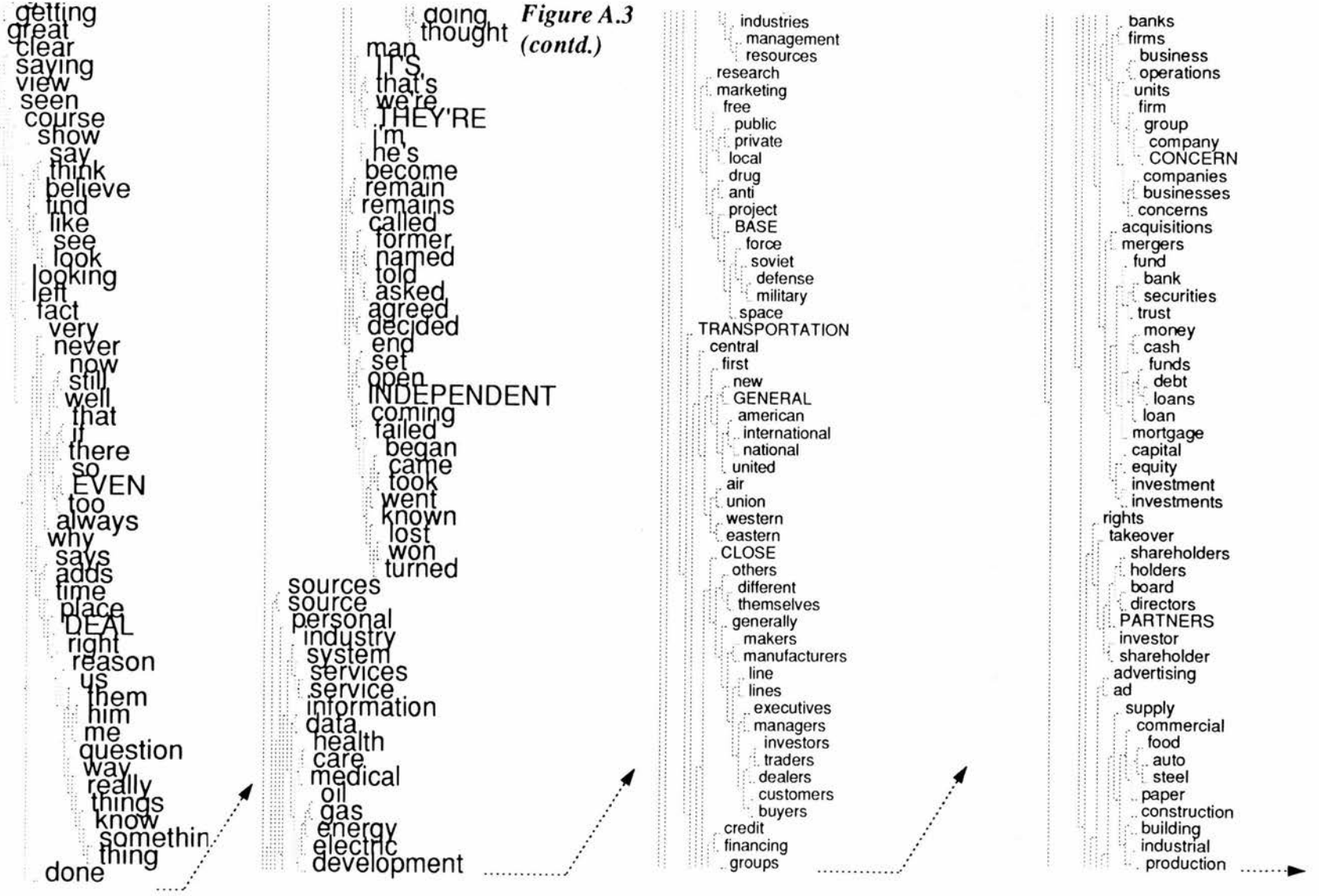


Figure A.3 (contd.)

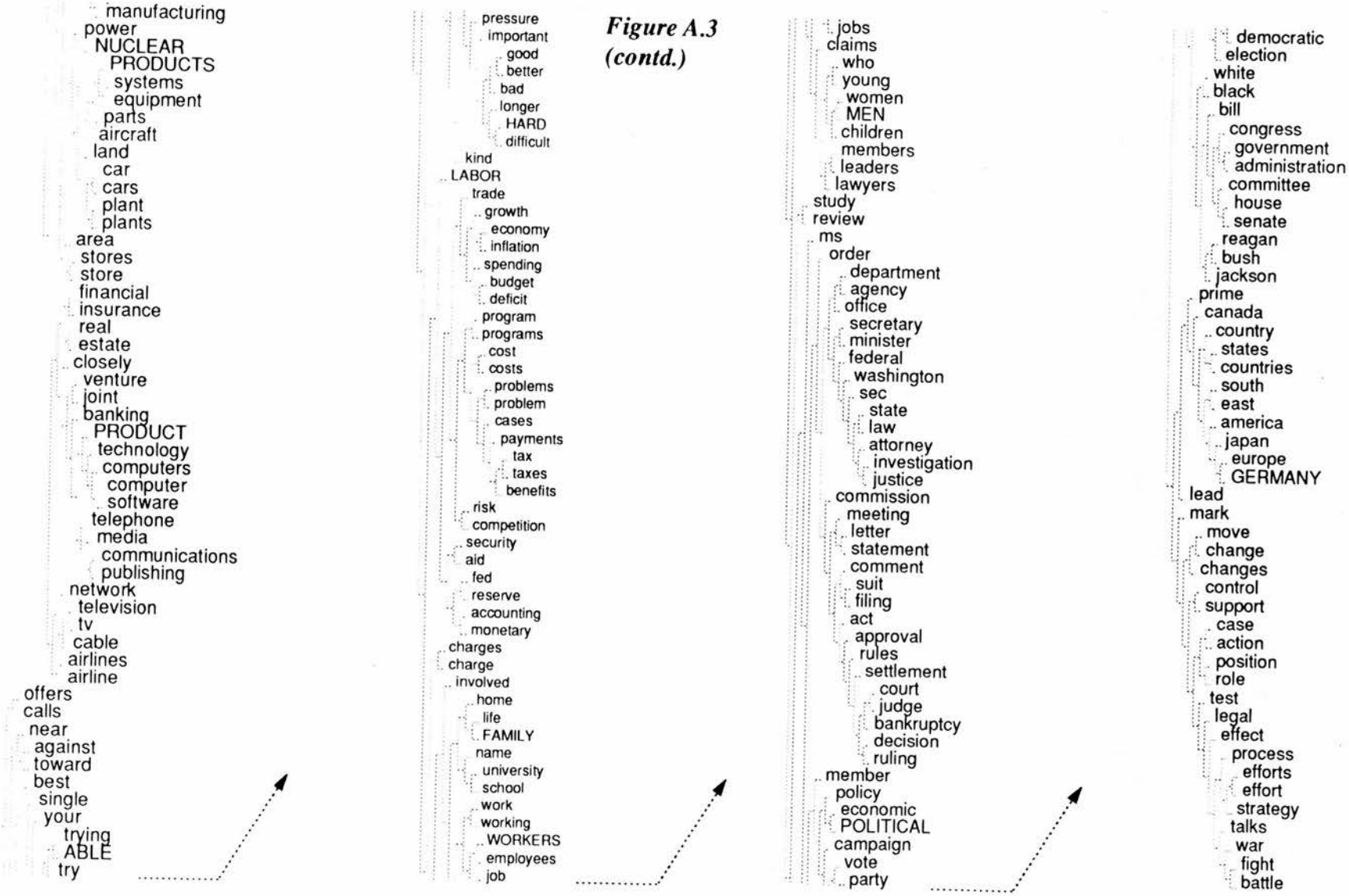


Figure A.3 (contd.)

Table A.4
(Euclidean Distance Metric, Window Length=2)

The table below contains the 50 target words and 10 nearest neighbours for analysis 4 in Chapter 4.

Target Word	10 Nearest Neighbours (Euclidean Distance)
able	try (0.148) trying (0.154) continue (0.155) going (0.161) likely (0.170) acquire (0.170) according (0.174) agreed (0.175) seems (0.177) want (0.180)
above	before (0.066) below (0.067) of (0.067) on (0.071) under (0.071) despite (0.071) during (0.072) system (0.074) around (0.074) project (0.074)
analyst	estimated (0.128) investor (0.135) official (0.137) independent (0.139) shearson (0.139) drexel (0.139) ibm (0.143) investment (0.145) additional (0.146) attorney (0.146)
base	tax (0.058) performance (0.061) price (0.061) strategy (0.062) job (0.063) public (0.063) division (0.063) huge (0.064) network (0.064) test (0.065)
close	yield (0.063) keep (0.089) lead (0.093) sell (0.094) failed (0.095) meet (0.098) take (0.099) efforts (0.099) gained (0.101) help (0.102)
concern	however (0.059) although (0.060) while (0.060) firm (0.060) that (0.062) in (0.062) also (0.064) fund (0.064) british (0.064) system (0.064)
deal	merger (0.069) agreement (0.071) proposal (0.076) test (0.079) man (0.079) day (0.080) move (0.080) line (0.080) problem (0.081) way (0.081)
despite	in (0.031) on (0.031) after (0.034) of (0.034) for (0.036) against (0.037) into (0.038) while (0.039) through (0.039) before (0.039)
even	but (0.047) while (0.049) today (0.049) that (0.050) without (0.052) to (0.052) and (0.053) when (0.053) with (0.054) by (0.054)
expects	plans (0.106) wants (0.113) decided (0.122) agreed (0.124) seems (0.128) make (0.134) failed (0.138) take (0.141) help (0.142) keep (0.142)
family	performance (0.043) state (0.044) public (0.047) price (0.047) french (0.048) job (0.048) office (0.049) democratic (0.050) division (0.050) campaign (0.051)
gained	failed (0.052) keep (0.059) fell (0.061) efforts (0.064) go (0.065) sell (0.067) rose (0.067) plans (0.069) dropped (0.069) raise (0.070)
george	robert (0.066) michael (0.066) john (0.068) richard (0.069) james (0.069) william (0.069) david (0.070) paul (0.073) sen (0.073) rep (0.074)
germany	german (0.100) canada (0.156) japan (0.160) south (0.160) hong (0.162) manufacturing (0.162) investments (0.164) both (0.164) western (0.164) california (0.165)
general	power (0.056) western (0.058) management (0.058) steel (0.058) credit (0.059) black (0.060) chemical (0.060) american (0.061) non (0.061) data (0.061)
hard	hard (0.000) him (0.058) me (0.059) them (0.062) continued (0.063) hold (0.066) dropped (0.066) enough (0.066) related (0.068) cut (0.068) get (0.068)
included	won (0.069) from (0.075) with (0.075) for (0.076) after (0.076) includes (0.078) issued (0.078) including (0.079) washington (0.079) big (0.080)
independent	investor (0.067) inflation (0.067) additional (0.067) investment (0.070) increasing (0.070) employees (0.071) american (0.071) economic (0.071) all (0.072) outside (0.072)
index	american (0.075) while (0.080) on (0.080) in (0.081) meanwhile (0.082) prices (0.082) funds (0.083) for (0.084) fund (0.084) and (0.085)
it's	he's (0.035) got (0.048) like (0.049) that's (0.052) they're (0.053) just (0.054) is (0.055) having (0.056) without (0.057) very (0.057)
labor	defense (0.045) british (0.054) while (0.055) in (0.056) meanwhile (0.056) and (0.056) european (0.056) military (0.057) black (0.057) for (0.057)
making	using (0.043) made (0.046) to (0.047) today (0.047) without (0.047) just (0.047) getting (0.047) and (0.048) with (0.048) where (0.048)
men	leaders (0.058) companies (0.061) problems (0.063) children (0.064) groups (0.064) military (0.065) lawyers (0.065) european (0.066) japanese (0.066) left (0.067)
night	month (0.070) week (0.075) fall (0.076) friday (0.078) after (0.081) until (0.082) while (0.082) wednesday (0.083) monday (0.083) began (0.084)
nuclear	power (0.067) and (0.067) non (0.068) credit (0.071) plants (0.071) black (0.071) partners (0.071) military (0.072) south (0.072) public (0.072)
old	five (0.082) last (0.089) ago (0.097) full (0.099) seven (0.100) later (0.102) three (0.103) period (0.103) four (0.103) this (0.106)
operating	financial (0.092) chairman (0.094) both (0.098) production (0.098) manufacturing (0.099) construction (0.099) marketing (0.102) management (0.102) steel (0.102) economic (0.103)

Table A.4 (contd.)

paid	sold (0.048) offered (0.053) made (0.055) today (0.056) further (0.058) using (0.062) itself (0.063) making (0.063) only (0.064) was (0.064)
partners	and (0.043) non (0.044) management (0.046) fund (0.046) meanwhile (0.046) using (0.046) today (0.047) programs (0.047) black (0.048) with (0.048)
percentage	view (0.100) leading (0.107) total (0.109) one (0.110) risk (0.110) unit (0.110) growing (0.111) range (0.111) cost (0.111) great (0.111)
political	black (0.034) military (0.035) management (0.036) private (0.038) building (0.038) local (0.039) legal (0.040) non (0.043) cars (0.043) construction (0.043)
preferred	common (0.115) options (0.117) american (0.125) index (0.125) record (0.132) cash (0.132) prices (0.133) buying (0.134) debt (0.134) selling (0.134)
product	building (0.045) network (0.047) system (0.049) strategy (0.050) public (0.051) program (0.051) in (0.052) job (0.052) for (0.052) through (0.053)
products	businesses (0.040) both (0.043) steel (0.043) technology (0.043) equipment (0.045) construction (0.045) aircraft (0.047) other (0.048) energy (0.048) programs (0.049)
same	company's (0.089) latest (0.104) economy (0.110) nation's (0.117) dollar (0.117) biggest (0.122) transaction (0.123) company (0.126) previous (0.129) sec (0.130)
should	must (0.024) may (0.036) will (0.037) might (0.041) won't (0.042) could (0.042) would (0.043) can (0.048) wouldn't (0.069) can't (0.075)
take	make (0.039) get (0.040) provide (0.047) go (0.050) hold (0.052) give (0.052) help (0.054) find (0.054) keep (0.057) lead (0.058)
they're	we're (0.038) he's (0.043) are (0.045) i'm (0.052) were (0.053) it's (0.053) aren't (0.053) just (0.054) her (0.055) your (0.057)
times	while (0.061) in (0.061) meanwhile (0.062) for (0.063) on (0.064) today (0.065) problems (0.065) but (0.065) after (0.066) city (0.066)
transaction	economy (0.054) process (0.057) problem (0.058) company (0.060) dollar (0.062) project (0.064) study (0.068) ruling (0.068) case (0.069) issue (0.071)
transportation	defense (0.061) energy (0.062) construction (0.070) aircraft (0.071) steel (0.071) labor (0.072) products (0.073) both (0.073) management (0.073) technology (0.074)
use	support (0.049) planned (0.053) approval (0.056) aid (0.057) act (0.059) review (0.059) control (0.059) purchase (0.060) shareholders (0.062) start (0.062)
using	through (0.039) without (0.039) for (0.040) and (0.040) while (0.040) by (0.040) to (0.041) with (0.041) in (0.041) today (0.041)
wall	firms (0.364) news (0.366) labor (0.367) europ (0.368) taxes (0.369) on (0.370) gm (0.370) and (0.370) division (0.371) issues (0.371)
wednesday	tuesday (0.029) monday (0.044) friday (0.053) while (0.059) feb (0.059) profits (0.060) today (0.061) data (0.062) began (0.062) jan (0.063)
week	month (0.040) year (0.073) night (0.075) fall (0.077) year's (0.094) after (0.101) later (0.101) until (0.101) began (0.105) when (0.105)
weren't	wasn't (0.113) were (0.113) are (0.113) aren't (0.113) and (0.117) while (0.117) but (0.117) that (0.119) meanwhile (0.119) today (0.119)
will	would (0.029) could (0.030) won't (0.030) must (0.031) may (0.032) should (0.037) might (0.041) can (0.043) wouldn't (0.054) to (0.063)
william	john (0.020) robert (0.023) richard (0.025) david (0.025) michael (0.026) james (0.026) paul (0.037) sen (0.042) rep (0.043) mr (0.045)
workers	employees (0.031) programs (0.041) both (0.043) cars (0.043) costs (0.044) businesses (0.044) production (0.045) lawyers (0.045) jobs (0.045) people (0.045)

Figure A.4 below shows the dendrogram containing the 1000 target words considered in analysis 4 of Chapter 4.

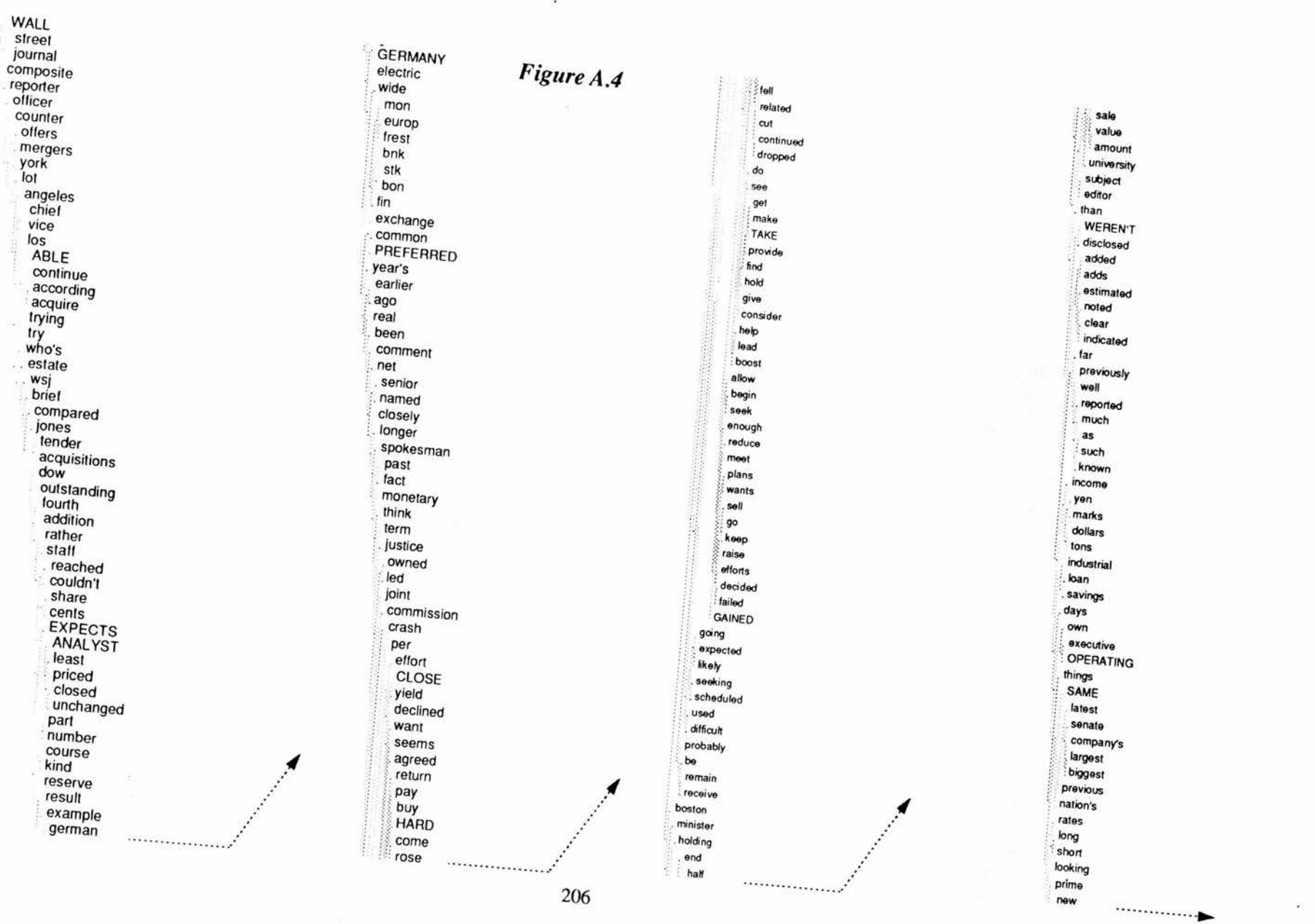

WALL
street
journal
composite
reporter
officer
counter
offers
mergers
york
lot
angeles
chief
vice
los
ABLE
continue
according
acquire
trying
try
who's
estate
wsj
brief
compared
jones
tender
acquisitions
dow
outstanding
fourth
addition
rather
staff
reached
couldn't
share
cents
EXPECTS
ANALYST
least
priced
closed
unchanged
part
number
course
kind
reserve
result
example
german
GERMANY
electric
wide
mon
europ
frest
bnk
stk
bon
fin
exchange
common
PREFERRED
year's
earlier
ago
real
been
comment
net
senior
named
closely
longer
spokesman
past
fact
monetary
think
term
justice
owned
led
joint
commission
crash
per
effort
CLOSE
yield
declined
want
seems
agreed
return
pay
buy
HARD
come
rose
fell
related
cut
continued
dropped
do
see
get
make
TAKE
provide
find
hold
give
consider
help
lead
boost
allow
begin
seek
enough
reduce
meet
plans
wants
sell
go
keep
raise
efforts
decided
failed
GAINED
going
expected
likely
seeking
scheduled
used
difficult
probably
be
remain
receive
boston
minister
holding
end
half
sale
value
amount
university
subject
editor
than
WEREN'T
disclosed
added
adds
estimated
noted
clear
indicated
far
previously
well
reported
much
as
such
known
income
yen
marks
dollars
tons
industrial
loan
savings
days
own
executive
OPERATING
things
SAME
latest
senate
company's
largest
biggest
previous
nation's
rates
long
short
looking
prime
new

Figure A.4

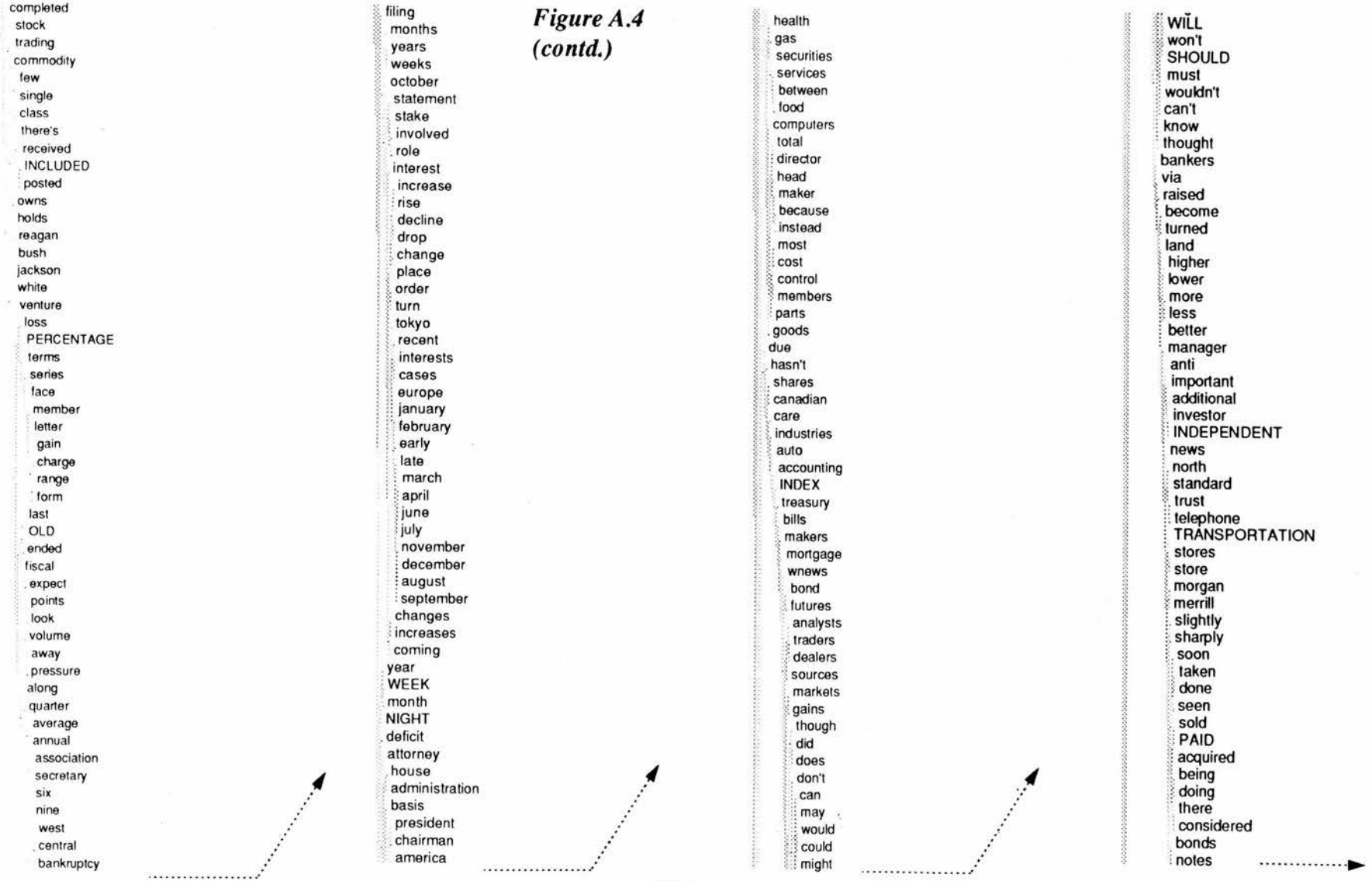


Figure A.4 (contd.)

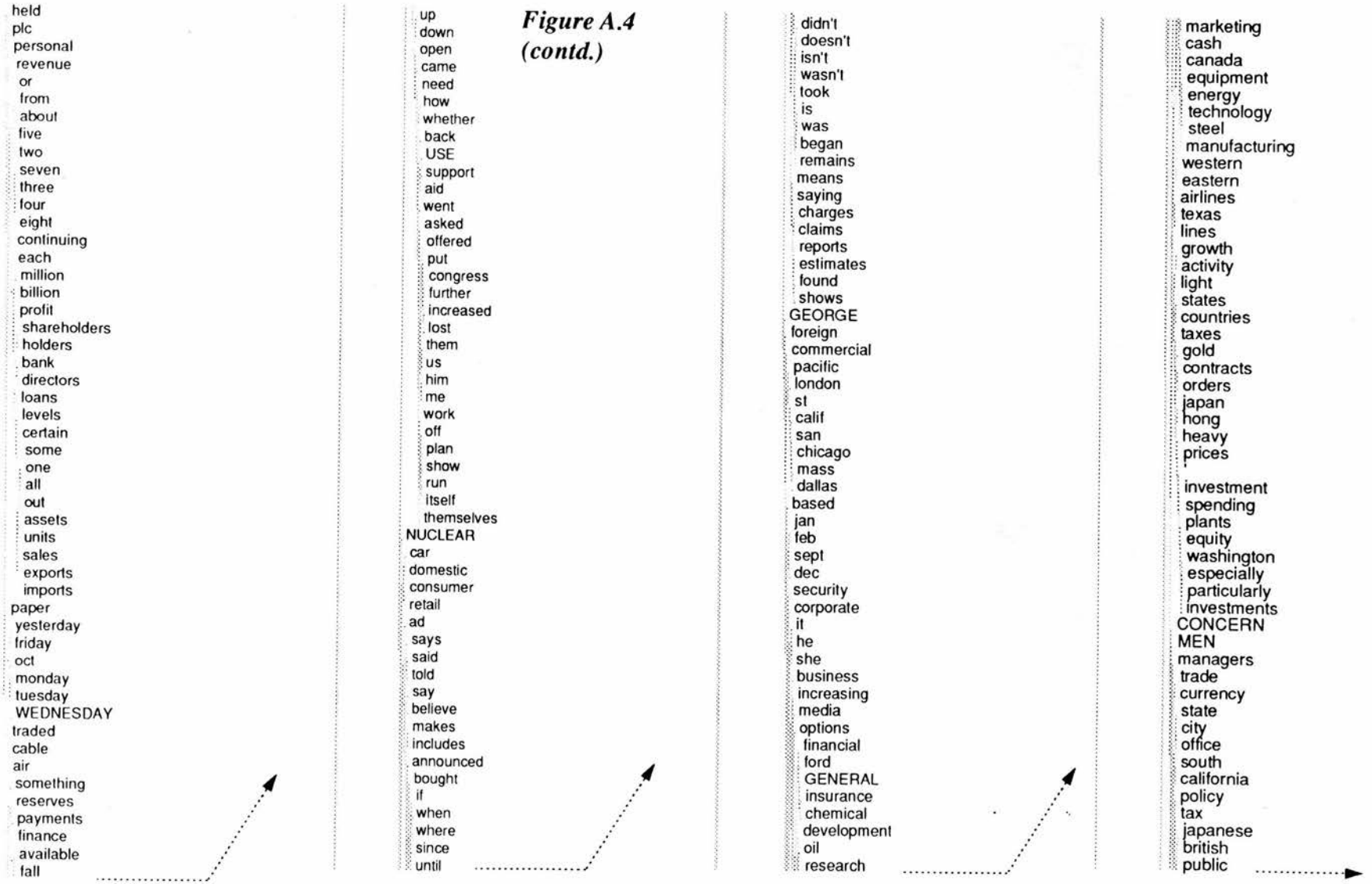


Figure A.4 (contd.)

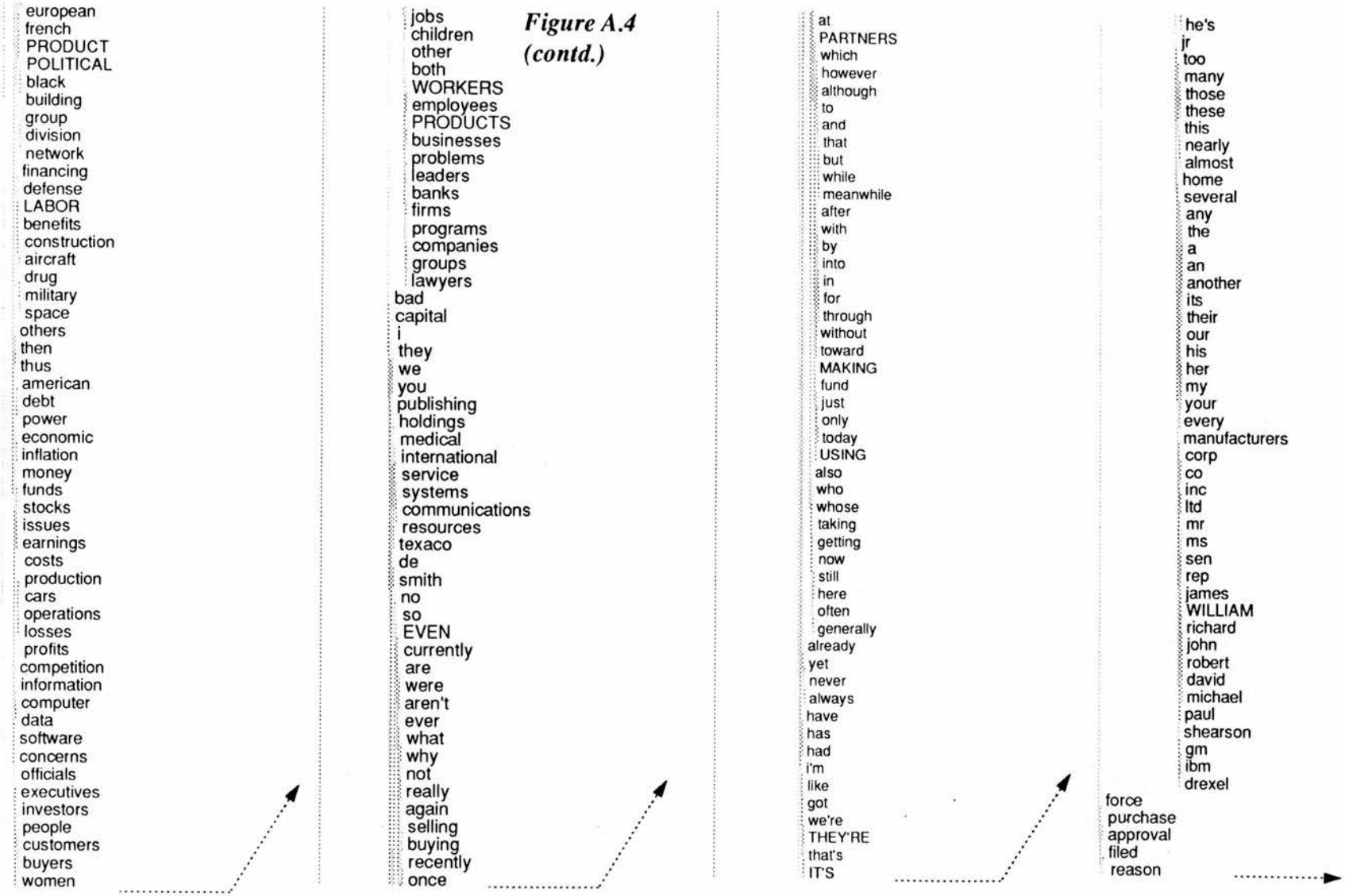


Figure A.4 (contd.)

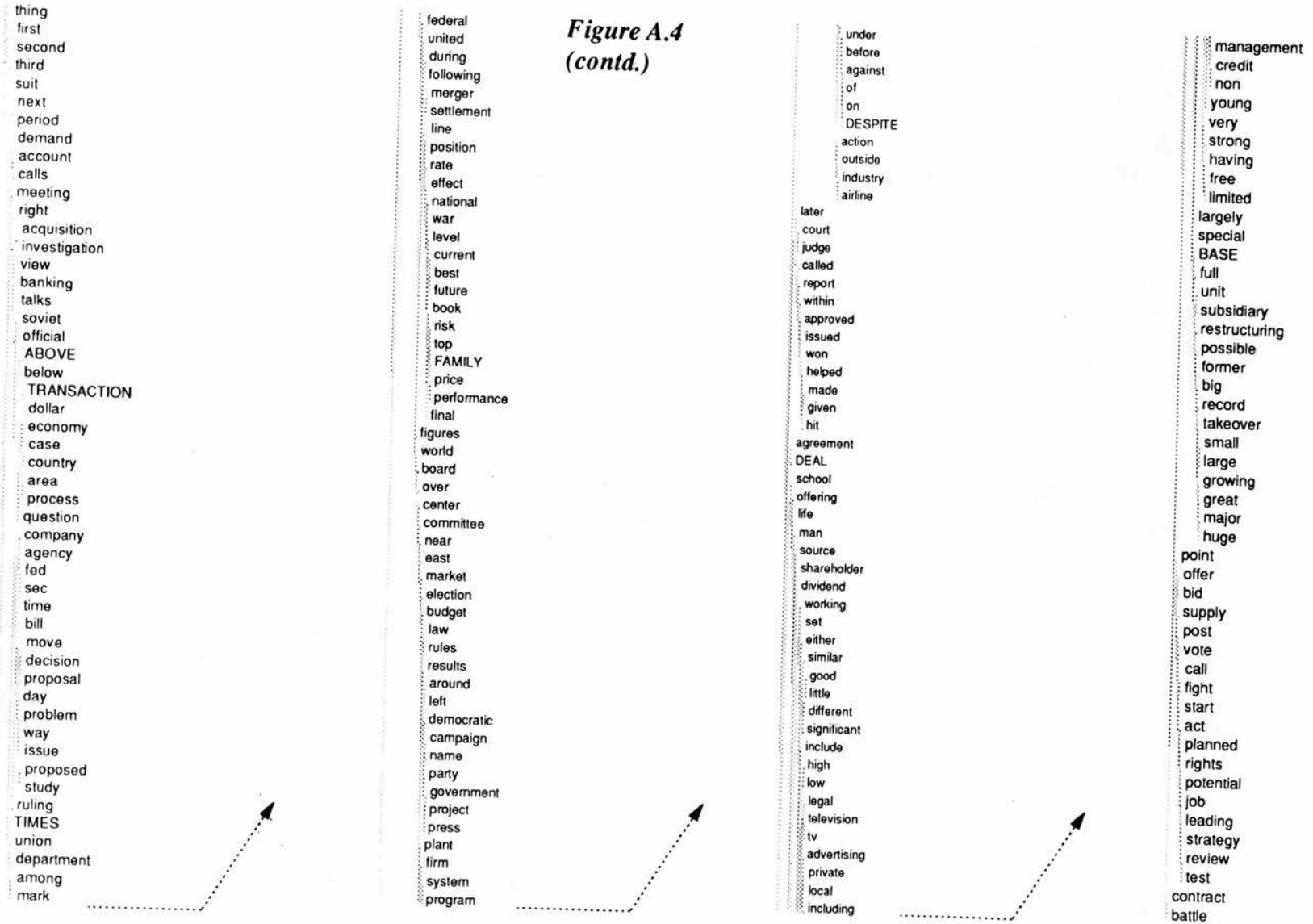


Figure A.4 (contd.)

Table A.5
(Spearman Distance Metric, Window Length=5)

The table below contains the 50 target words and 10 nearest neighbours for analysis 5 in Chapter 4.

Target Word	10 Nearest Neighbours (Spearman Rank Correlation Coefficient)
able	can (0.734) if (0.702) they (0.697) do (0.693) can't (0.693) even (0.689) any (0.688) not (0.688) want (0.687) get (0.687)
above	below (0.745) higher (0.684) average (0.683) than (0.682) down (0.677) lower (0.672) level (0.669) price (0.668) up (0.660) year (0.659)
analyst	at (0.606) analysts (0.604) said (0.591) smith (0.567) and (0.556) added (0.549) big (0.549) according (0.549) chief (0.543) market (0.543)
base	and (0.615) cost (0.611) more (0.603) costs (0.597) their (0.594) most (0.592) the (0.592) for (0.592) well (0.591) large (0.590)
close	at (0.696) down (0.691) closed (0.687) friday (0.680) up (0.679) end (0.679) after (0.676) the (0.676) early (0.673) last (0.662)
concern	company (0.800) based (0.744) group (0.744) unit (0.721) said (0.713) inc (0.709) maker (0.709) corp (0.694) holding (0.693) services (0.685)
deal	out (0.688) he (0.674) they (0.664) be (0.660) it (0.660) so (0.660) what (0.660) think (0.659) that (0.658) say (0.658)
despite	while (0.714) result (0.699) in (0.699) continued (0.690) on (0.689) the (0.688) as (0.684) recent (0.683) because (0.683) however (0.682)
even	but (0.862) so (0.850) though (0.818) they (0.817) that (0.810) if (0.809) still (0.809) some (0.806) many (0.805) much (0.802)
expects	expected (0.703) its (0.664) will (0.643) plans (0.641) quarter (0.635) net (0.634) reported (0.633) said (0.633) third (0.621) earnings (0.617)
family	his (0.634) mr (0.632) who (0.627) life (0.627) home (0.625) one (0.621) he (0.619) old (0.611) state (0.600) school (0.598)
gained	fell (0.741) rose (0.740) dropped (0.704) closed (0.686) trading (0.657) posted (0.653) volume (0.641) ended (0.638) reported (0.636) share (0.636)
george	his (0.644) president (0.638) john (0.632) mr (0.626) james (0.622) old (0.621) robert (0.620) paul (0.619) says (0.617) bush (0.614)
germany	west (0.704) japan (0.656) german (0.637) europe (0.633) european (0.608) japanese (0.592) east (0.589) foreign (0.569) soviet (0.569) government (0.566)
general	of (0.675) group (0.673) and (0.672) co (0.671) new (0.670) a (0.666) said (0.660) corp (0.658) national (0.658) american (0.656)
hard	so (0.739) like (0.737) it's (0.733) do (0.723) way (0.719) too (0.718) not (0.715) them (0.712) people (0.711) think (0.710)
included	include (0.656) by (0.635) million (0.632) including (0.630) of (0.629) from (0.623) includes (0.622) for (0.615) reported (0.615) which (0.614)
independent	members (0.590) public (0.584) agency (0.582) state (0.581) own (0.580) outside (0.563) mr (0.562) management (0.561) national (0.555) information (0.555)
index	prices (0.732) stocks (0.725) traders (0.717) trading (0.717) futures (0.703) volume (0.698) stock (0.693) market (0.686) average (0.685) price (0.682)
it's	you (0.846) we (0.832) not (0.826) says (0.826) think (0.826) i (0.821) do (0.820) so (0.817) going (0.812) what (0.809)
labor	washington (0.581) department (0.577) trade (0.577) state (0.573) federal (0.571) workers (0.571) force (0.568) health (0.566) economic (0.565) and (0.563)
making	make (0.762) made (0.730) so (0.709) such (0.707) as (0.702) all (0.695) more (0.692) because (0.690) some (0.690) but (0.689)
men	women (0.712) people (0.705) who (0.671) him (0.667) she (0.660) her (0.657) i (0.650) white (0.645) them (0.643) my (0.641)
night	his (0.641) mr (0.634) him (0.610) he (0.608) here (0.596) when (0.592) old (0.591) until (0.591) black (0.587) she (0.585)
nuclear	plant (0.640) power (0.630) plants (0.608) defense (0.599) soviet (0.589) military (0.583) force (0.574) project (0.570) government (0.561) work (0.554)
old	who (0.744) his (0.711) says (0.706) mr (0.698) her (0.678) he (0.675) i (0.672) him (0.671) years (0.668) she (0.659)
operating	operations (0.744) profit (0.707) sales (0.705) earnings (0.703) net (0.690) quarter (0.681) company's (0.677) revenue (0.672) company (0.662) fourth (0.660)

Table A.5 (contd.)

paid	pay (0.737) for (0.666) or (0.664) about (0.657) of (0.641) from (0.639) as (0.635) each (0.633) than (0.631) the (0.629)
partners	group (0.679) firm (0.657) company (0.642) management (0.626) investment (0.622) inc (0.601) co (0.597) acquisition (0.594) shareholders (0.591) holding (0.589)
percentage	rate (0.686) average (0.684) year (0.680) compared (0.667) than (0.663) slightly (0.648) total (0.645) higher (0.643) points (0.639) half (0.633)
political	democratic (0.749) policy (0.708) party (0.700) campaign (0.686) reagan (0.676) bush (0.673) war (0.667) congress (0.664) what (0.661) his (0.658)
preferred	common (0.722) holders (0.703) shares (0.701) outstanding (0.672) stock (0.667) purchase (0.654) transaction (0.654) offer (0.651) million (0.650) share (0.649)
product	products (0.662) line (0.629) computer (0.623) wide (0.618) industry (0.611) computers (0.611) sales (0.611) business (0.610) customers (0.610) this (0.609)
products	equipment (0.728) business (0.712) systems (0.709) technology (0.707) services (0.703) computer (0.695) computers (0.690) businesses (0.686) parts (0.684) food (0.676)
same	only (0.744) that (0.740) but (0.729) even (0.724) than (0.723) more (0.723) all (0.719) as (0.704) one (0.702) most (0.702)
should	would (0.811) could (0.801) can (0.794) not (0.794) must (0.784) if (0.781) might (0.772) will (0.754) they (0.754) don't (0.739)
take	make (0.751) go (0.750) get (0.746) them (0.745) they (0.744) that (0.744) do (0.743) but (0.739) have (0.738) if (0.738)
they're	it's (0.770) you (0.753) do (0.741) so (0.740) like (0.736) people (0.733) think (0.732) not (0.731) get (0.731) them (0.729)
times	years (0.667) one (0.654) time (0.650) around (0.647) than (0.646) at (0.642) past (0.642) just (0.640) the (0.638) about (0.638)
transaction	acquisition (0.723) offer (0.712) purchase (0.693) sale (0.687) company (0.679) merger (0.669) holders (0.665) shareholders (0.661) preferred (0.654) agreement (0.647)
transportation	department (0.595) services (0.593) energy (0.579) general (0.565) international (0.565) corp (0.554) united (0.554) co (0.553) air (0.551) new (0.546)
use	used (0.738) such (0.734) using (0.732) make (0.690) provide (0.682) have (0.678) can (0.677) their (0.675) own (0.675) other (0.675)
using	use (0.732) used (0.703) such (0.689) can (0.679) into (0.654) more (0.652) other (0.652) their (0.648) own (0.648) and (0.644)
wall	street (0.947) journal (0.794) news (0.639) york (0.634) new (0.631) wsj (0.629) international (0.619) staff (0.605) unit (0.601) co (0.596)
wednesday	monday (0.755) friday (0.747) tuesday (0.709) yesterday (0.709) week (0.703) late (0.697) last (0.663) month (0.662) june (0.642) april (0.635)
week	month (0.806) last (0.795) friday (0.763) yesterday (0.744) late (0.731) april (0.723) monday (0.723) year (0.723) after (0.720) the (0.717)
weren't	terms (0.655) disclosed (0.635) couldn't (0.621) had (0.621) wasn't (0.620) said (0.620) two (0.617) were (0.614) also (0.608) been (0.607)
will	would (0.863) may (0.801) could (0.790) to (0.776) won't (0.773) should (0.754) be (0.730) next (0.719) is (0.717) can (0.714)
william	robert (0.695) john (0.694) james (0.678) david (0.673) chairman (0.668) chief (0.654) president (0.652) richard (0.648) executive (0.642) jr (0.641)
workers	employees (0.723) work (0.693) jobs (0.681) plant (0.635) members (0.628) women (0.625) people (0.619) those (0.619) states (0.616) state (0.615)

Figure A.5 below shows the dendrogram containing the 1000 target words considered in analysis 5 of Chapter 4.

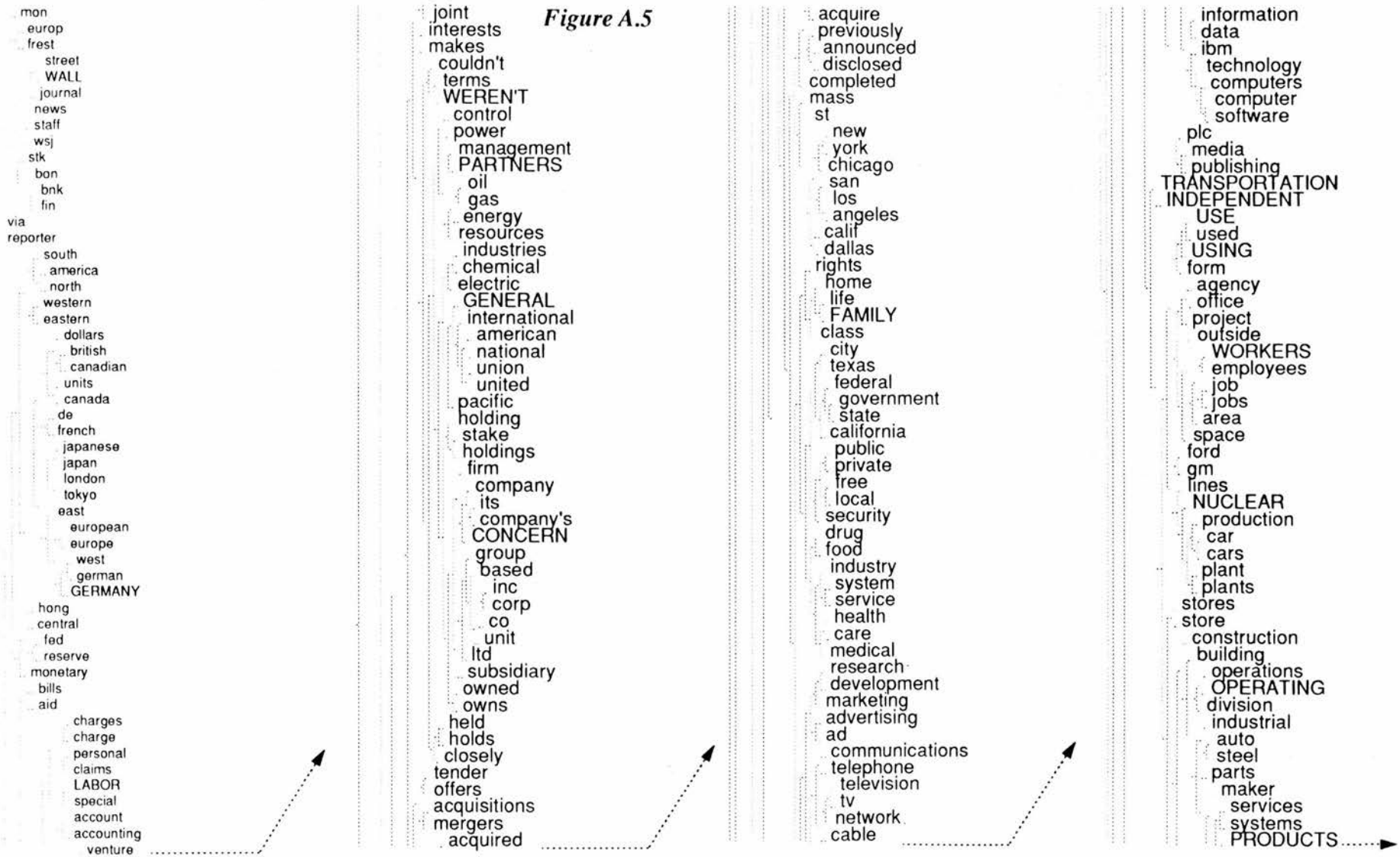

mon
europ
frest
street
WALL
journal
news
staff
wsj
stk
bon
bnk
fin
via
reporter
south
america
north
western
eastern
dollars
british
canadian
units
canada
de
french
japanese
japan
london
tokyo
east
european
europe
west
german
GERMANY
hong
central
fed
reserve
monetary
bills
aid
charges
charge
personal
claims
LABOR
special
account
accounting
venture
joint
interests
makes
couldn't
terms
WEREN'T
control
power
management
PARTNERS
oil
gas
energy
resources
industries
chemical
electric
GENERAL
international
american
national
union
united
pacific
holding
stake
holdings
firm
company
its
company's
CONCERN
group
based
inc
corp
co
unit
ltd
subsidiary
owned
owns
held
holds
closely
tender
offers
acquisitions
mergers
acquired
acquire
previously
announced
disclosed
completed
mass
st
new
york
chicago
san
los
angeles
calif
dallas
rights
home
life
FAMILY
class
city
texas
federal
government
state
california
public
private
free
local
security
drug
food
industry
system
service
health
care
medical
research
development
marketing
advertising
ad
communications
telephone
television
tv
network
cable
information
data
ibm
technology
computers
computer
software
plc
media
publishing
TRANSPORTATION
INDEPENDENT
USE
used
USING
form
agency
office
project
outside
WORKERS
employees
job
jobs
area
space
ford
gm
lines
NUCLEAR
production
car
cars
plant
plants
stores
store
construction
building
operations
OPERATING
division
industrial
auto
steel
parts
maker
services
systems
PRODUCTS

Figure A.5

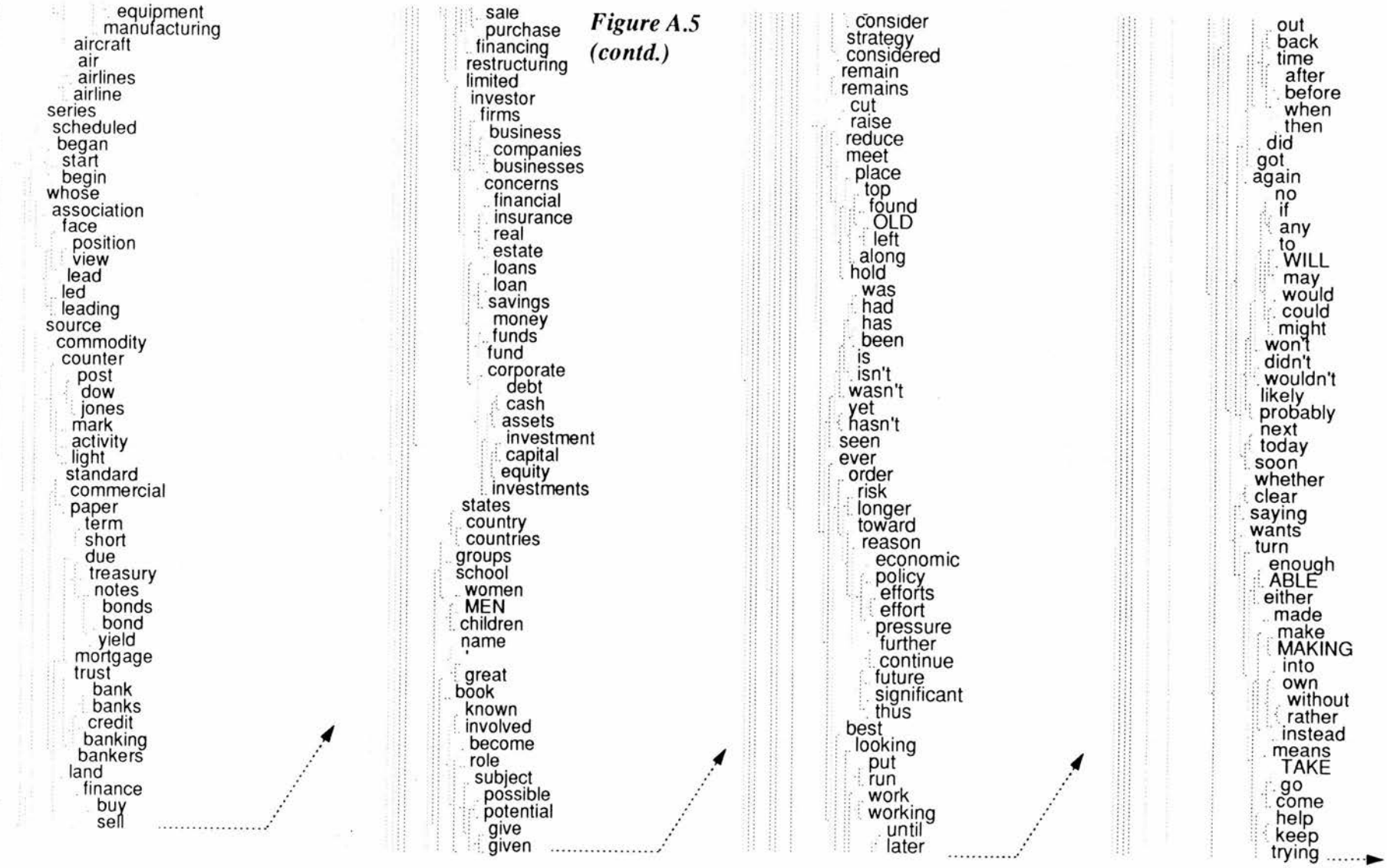


Figure A.5 (contd.)

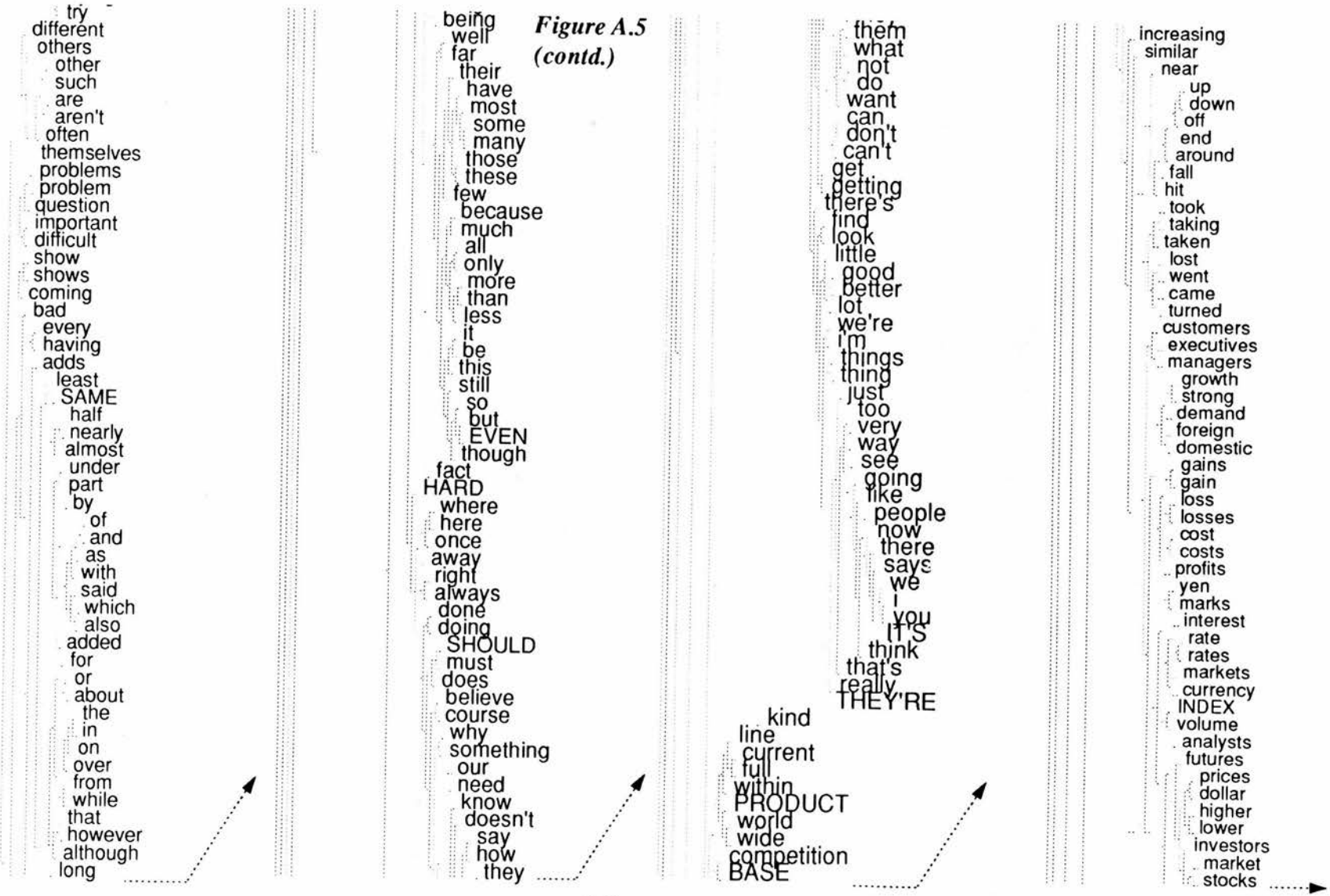


Figure A.5 (contd.)

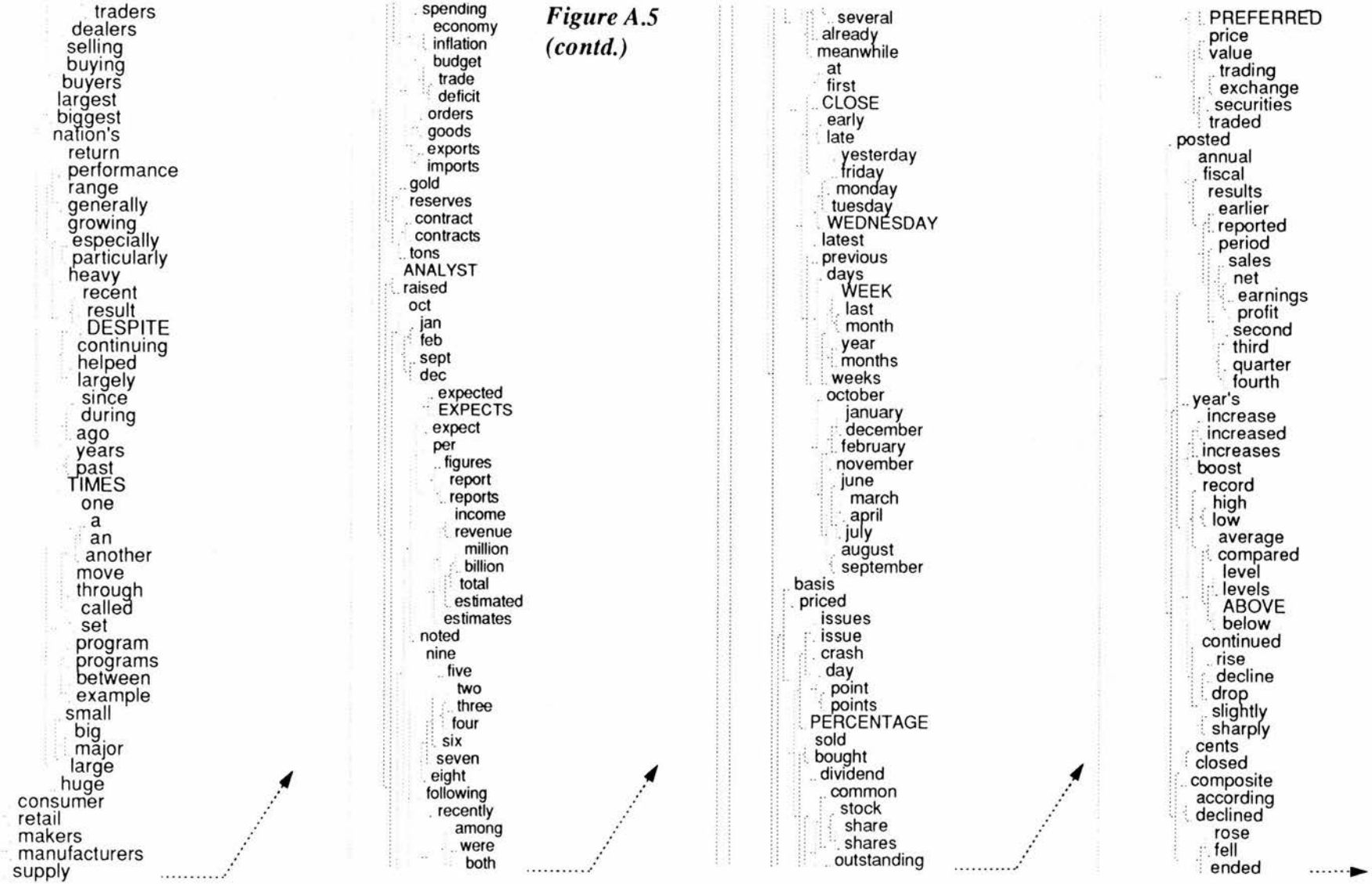


Figure A.5 (contd.)

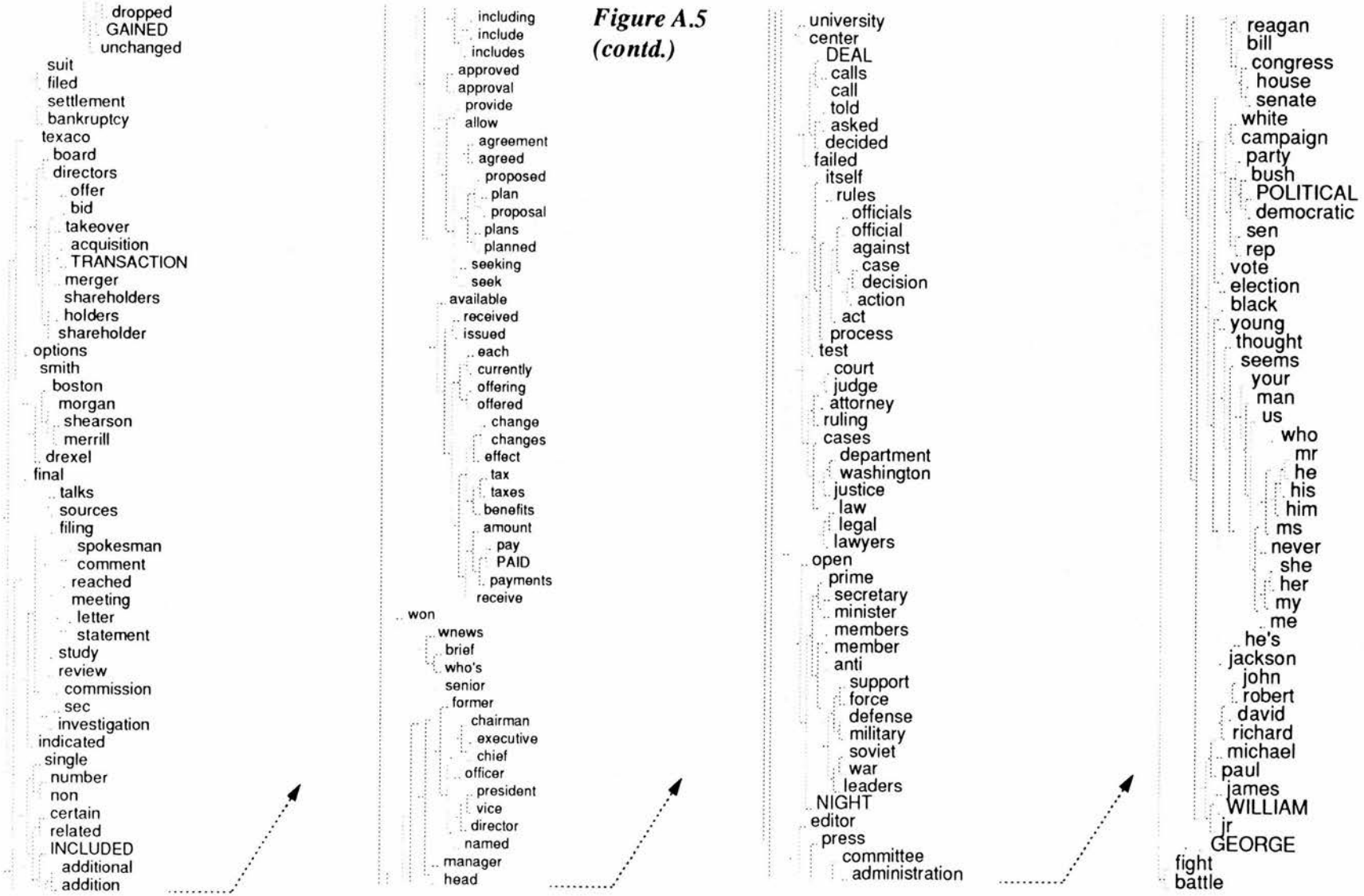


Figure A.5 (contd.)

Table A.6
(Euclidean Distance Metric, Window Length=5)

The table below contains the 50 target words and 10 nearest neighbours for analysis 6 in Chapter 4.

Target Word	10 Nearest Neighbours (Euclidean Distance)
able	try (0.059) allow (0.061) going (0.062) continue (0.062) trying (0.063) likely (0.067) seems (0.067) enough (0.068) want (0.068) meet (0.068)
above	below (0.035) rate (0.039) during (0.039) issue (0.040) price (0.040) before (0.041) level (0.041) current (0.041) under (0.042) final (0.042)
analyst	smith (0.072) drexel (0.073) inc (0.073) merrill (0.075) co (0.076) official (0.076) calif (0.077) mr (0.077) paul (0.077) morgan (0.078)
base	line (0.031) product (0.032) huge (0.032) price (0.032) strategy (0.032) tax (0.032) major (0.033) public (0.033) performance (0.033) top (0.033)
close	dropped (0.045) take (0.046) keep (0.046) lead (0.046) failed (0.047) gained (0.049) start (0.049) help (0.049) back (0.049) decided (0.049)
concern	building (0.032) for (0.033) firm (0.033) in (0.033) energy (0.033) service (0.033) with (0.034) network (0.034) into (0.034) product (0.035)
deal	merger (0.032) agreement (0.034) move (0.035) way (0.035) line (0.036) show (0.036) test (0.036) problem (0.036) program (0.037) proposal (0.037)
despite	after (0.021) on (0.022) in (0.022) performance (0.022) of (0.023) against (0.023) before (0.024) through (0.025) public (0.025) airline (0.025)
even	but (0.021) just (0.022) to (0.022) that (0.023) generally (0.023) now (0.023) still (0.024) often (0.024) only (0.024) not (0.024)
expects	plans (0.048) agreed (0.054) wants (0.054) seek (0.057) decided (0.058) sell (0.059) make (0.063) expect (0.063) acquire (0.064) raise (0.064)
family	state (0.024) public (0.025) leading (0.026) top (0.027) performance (0.027) office (0.027) review (0.027) book (0.028) political (0.028) system (0.029)
gained	sell (0.044) failed (0.044) dropped (0.045) keep (0.045) efforts (0.045) fell (0.046) plans (0.046) meet (0.046) agreed (0.048) raise (0.049)
george	robert (0.028) james (0.029) richard (0.029) john (0.030) william (0.030) michael (0.031) david (0.032) paul (0.033) sen (0.035) former (0.036)
germany	german (0.052) canada (0.070) japan (0.070) south (0.072) hong (0.073) east (0.073) western (0.074) london (0.074) investments (0.074) manufacturing (0.075)
general	western (0.029) management (0.029) american (0.029) financial (0.029) chemical (0.029) data (0.029) power (0.029) energy (0.030) united (0.030) international (0.030)
hard	them (0.029) get (0.030) find (0.030) make (0.030) enough (0.031) him (0.031) go (0.032) help (0.032) see (0.033) take (0.033)
included	charge (0.041) about (0.042) year (0.042) includes (0.044) nearly (0.044) from (0.046) gain (0.046) issued (0.046) profit (0.046) after (0.046)
independent	outside (0.028) using (0.031) all (0.032) power (0.033) economic (0.033) through (0.033) public (0.033) almost (0.033) anti (0.033) certain (0.034)
index	market (0.048) price (0.051) on (0.052) currency (0.052) before (0.053) oct (0.053) current (0.053) points (0.054) off (0.054) despite (0.054)
it's	he's (0.025) like (0.026) that's (0.027) little (0.028) we're (0.028) got (0.028) good (0.028) just (0.029) having (0.029) very (0.030)
labor	by (0.029) defense (0.029) and (0.030) in (0.030) washington (0.031) through (0.031) to (0.031) for (0.032) with (0.032) on (0.032)
making	with (0.021) made (0.022) having (0.022) now (0.022) using (0.022) to (0.022) private (0.022) but (0.023) just (0.023) non (0.023)
men	leaders (0.031) groups (0.032) children (0.033) lawyers (0.034) problems (0.034) companies (0.034) women (0.034) people (0.034) military (0.035) political (0.035)
night	month (0.035) after (0.035) was (0.036) week (0.036) when (0.036) began (0.037) fall (0.037) in (0.037) made (0.038) took (0.038)
nuclear	power (0.028) plant (0.031) public (0.031) which (0.032) through (0.032) military (0.033) in (0.033) defense (0.033) european (0.033) on (0.034)
old	five (0.043) last (0.043) this (0.046) later (0.046) full (0.048) seven (0.049) who (0.049) three (0.049) four (0.049) for (0.049)
operating	operations (0.044) financial (0.044) production (0.046) marketing (0.048) management (0.048) chairman (0.048) manufacturing (0.049) finance (0.049) including (0.049) non (0.049)
paid	offered (0.030) made (0.030) making (0.031) sold (0.031) potential (0.032) public (0.032) tax (0.032) which (0.032) financing (0.033) given (0.033)

Table A.6 (contd.)

partners	management (0.024) investment (0.024) and (0.026) limited (0.026) non (0.026) with (0.027) the (0.028) a (0.028) advertising (0.028) western (0.029)
percentage	leading (0.054) view (0.054) half (0.054) top (0.055) full (0.055) one (0.055) level (0.055) seven (0.055) rate (0.056) day (0.056)
political	black (0.019) military (0.021) public (0.021) legal (0.022) great (0.022) defense (0.022) policy (0.023) with (0.023) then (0.023) through (0.023)
preferred	common (0.052) class (0.060) cash (0.062) each (0.062) holders (0.063) options (0.063) offering (0.063) including (0.064) shares (0.065) record (0.065)
product	building (0.024) huge (0.024) line (0.024) network (0.024) public (0.025) political (0.025) strategy (0.025) in (0.025) for (0.025) with (0.025)
products	equipment (0.023) businesses (0.025) medical (0.028) technology (0.028) construction (0.028) systems (0.029) marketing (0.029) lines (0.029) manufacturing (0.029) services (0.029)
same	company's (0.042) economy (0.044) latest (0.047) biggest (0.049) dollar (0.051) during (0.052) nation's (0.052) process (0.053) senate (0.054) best (0.054)
should	must (0.015) may (0.019) will (0.019) would (0.021) could (0.022) might (0.022) won't (0.024) can (0.025) wouldn't (0.034) can't (0.035)
take	make (0.021) get (0.022) give (0.022) help (0.023) go (0.024) lead (0.025) keep (0.026) find (0.027) hold (0.027) seek (0.028)
they're	we're (0.025) they (0.028) are (0.028) aren't (0.029) i'm (0.029) like (0.030) it's (0.030) their (0.030) just (0.030) often (0.030)
times	in (0.030) on (0.030) which (0.032) meanwhile (0.032) two (0.032) for (0.032) after (0.033) around (0.033) although (0.033) only (0.033)
transaction	company (0.029) project (0.030) process (0.033) study (0.035) final (0.035) economy (0.036) issue (0.037) agency (0.039) case (0.039) proposed (0.040)
transportation	energy (0.043) defense (0.045) industrial (0.047) financial (0.048) service (0.048) labor (0.048) medical (0.048) off (0.048) steel (0.048) construction (0.048)
use	support (0.024) aid (0.025) rights (0.026) work (0.027) help (0.028) hold (0.028) planned (0.029) certain (0.030) shareholders (0.030) them (0.031)
using	having (0.019) without (0.021) while (0.022) free (0.022) making (0.022) through (0.023) and (0.023) non (0.023) with (0.023) for (0.024)
wall	journal (0.127) street (0.133) reporter (0.140) europ (0.152) frest (0.158) news (0.165) labor (0.169) wnews (0.170) staff (0.171) japan (0.171)
wednesday	tuesday (0.018) monday (0.024) april (0.034) friday (0.035) march (0.035) up (0.037) while (0.037) heavy (0.037) late (0.038) june (0.039)
week	month (0.022) fall (0.035) night (0.036) year (0.038) year's (0.042) after (0.043) later (0.044) began (0.045) was (0.046) in (0.046)
weren't	were (0.051) are (0.052) also (0.053) in (0.053) officials (0.053) and (0.053) aren't (0.053) although (0.053) data (0.054) on (0.054)
will	would (0.017) may (0.018) must (0.018) should (0.019) could (0.020) won't (0.022) might (0.028) can (0.028) to (0.030) wouldn't (0.030)
william	robert (0.012) john (0.012) richard (0.014) james (0.015) david (0.016) michael (0.018) paul (0.021) jr (0.025) george (0.030) rep (0.031)
workers	employees (0.017) jobs (0.023) companies (0.023) programs (0.024) costs (0.024) leaders (0.024) local (0.024) groups (0.024) all (0.024) non (0.025)

Figure A.6 below shows the dendrogram containing the 1000 target words considered in analysis 6 of Chapter 4.

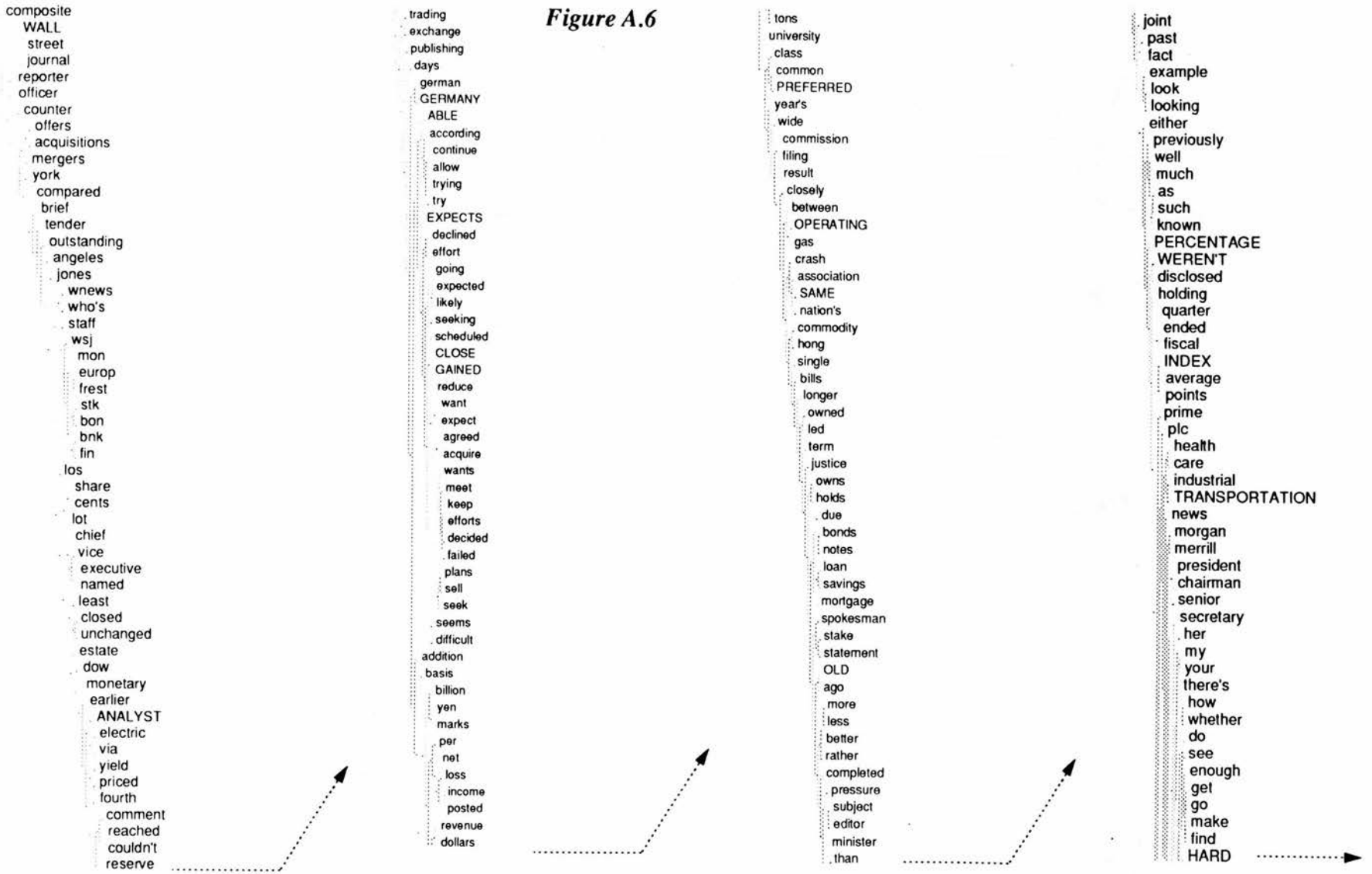

composite
WALL
street
journal
reporter
officer
counter
offers
acquisitions
mergers
york
compared
brief
tender
outstanding
angeles
jones
wnews
who's
staff
wsj
mon
europ
frest
stk
bon
bnk
fin
los
share
cents
lot
chief
vice
executive
named
least
closed
unchanged
estate
dow
monetary
earlier
ANALYST
electric
via
yield
priced
fourth
comment
reached
couldn't
reserve
trading
exchange
publishing
days
german
GERMANY
ABLE
according
continue
allow
trying
try
EXPECTS
declined
effort
going
expected
likely
seeking
scheduled
CLOSE
GAINED
reduce
want
expect
agreed
acquire
wants
meet
keep
efforts
decided
failed
plans
sell
seek
seems
difficult
addition
basis
billion
yen
marks
per
net
loss
income
posted
revenue
dollars
tons
university
class
common
PREFERRED
year's
wide
commission
filing
result
closely
between
OPERATING
gas
crash
association
SAME
nation's
commodity
hong
single
bills
longer
owned
led
term
justice
owns
holds
due
bonds
notes
loan
savings
mortgage
spokesman
stake
statement
OLD
ago
more
less
better
rather
completed
pressure
subject
editor
minister
than
joint
past
fact
example
look
looking
either
previously
well
much
as
such
known
PERCENTAGE
WEREN'T
disclosed
holding
quarter
ended
fiscal
INDEX
average
points
prime
plc
health
care
industrial
TRANSPORTATION
news
morgan
merrill
president
chairman
senior
secretary
her
my
your
there's
how
whether
do
see
enough
get
go
make
find
HARD

Figure A.6

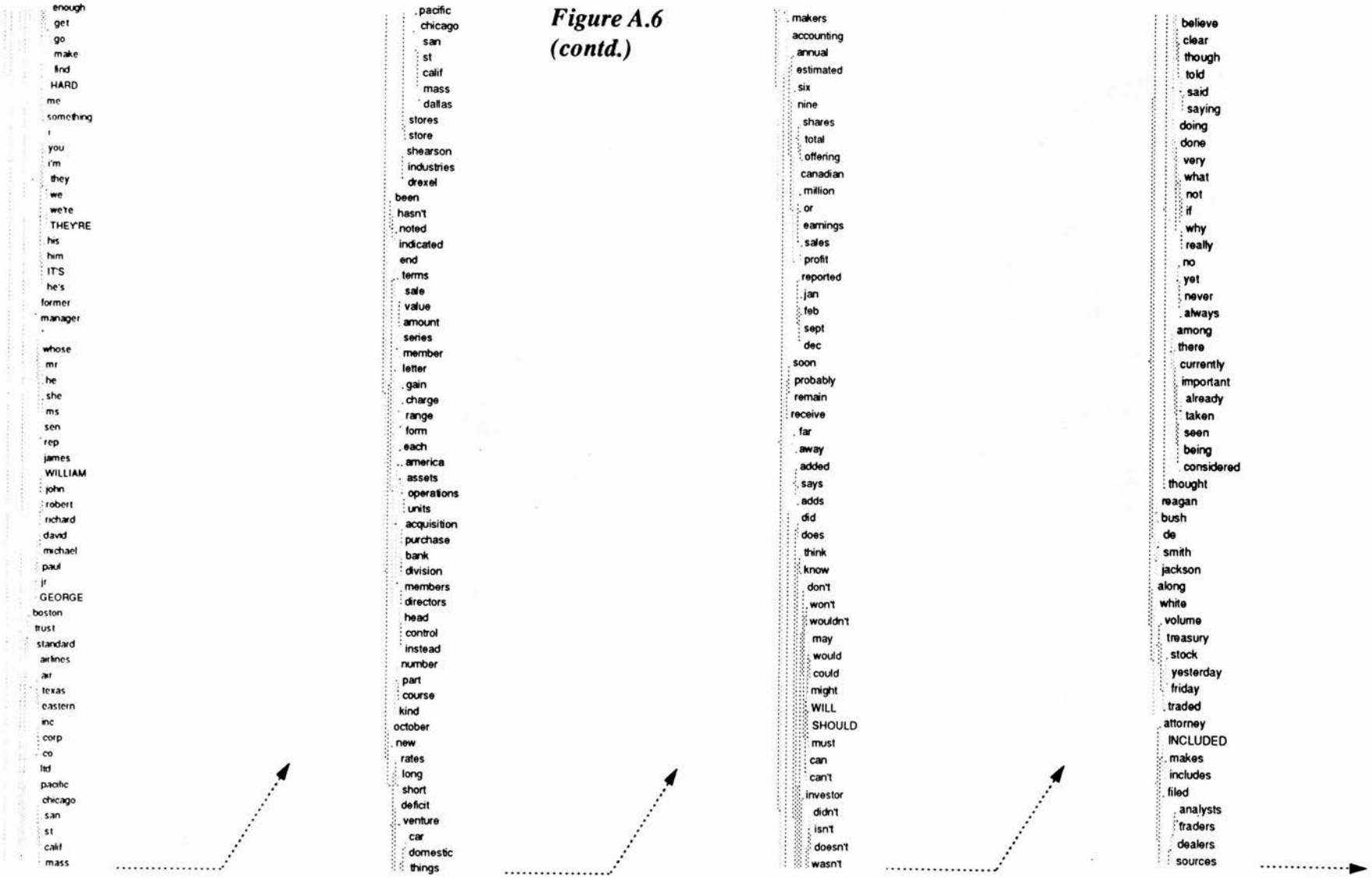


Figure A.6 (contd.)

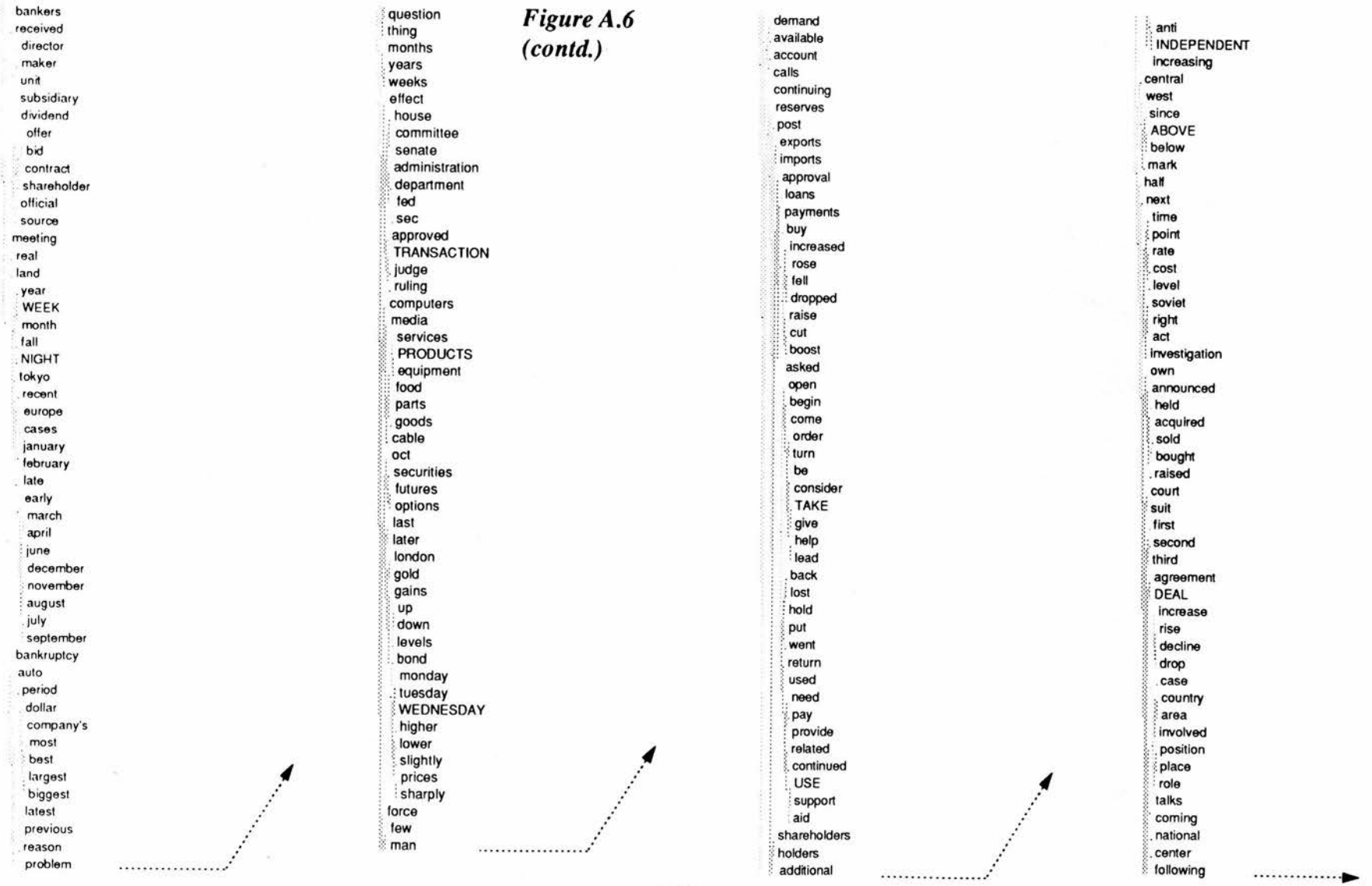


Figure A.6 (contd.)

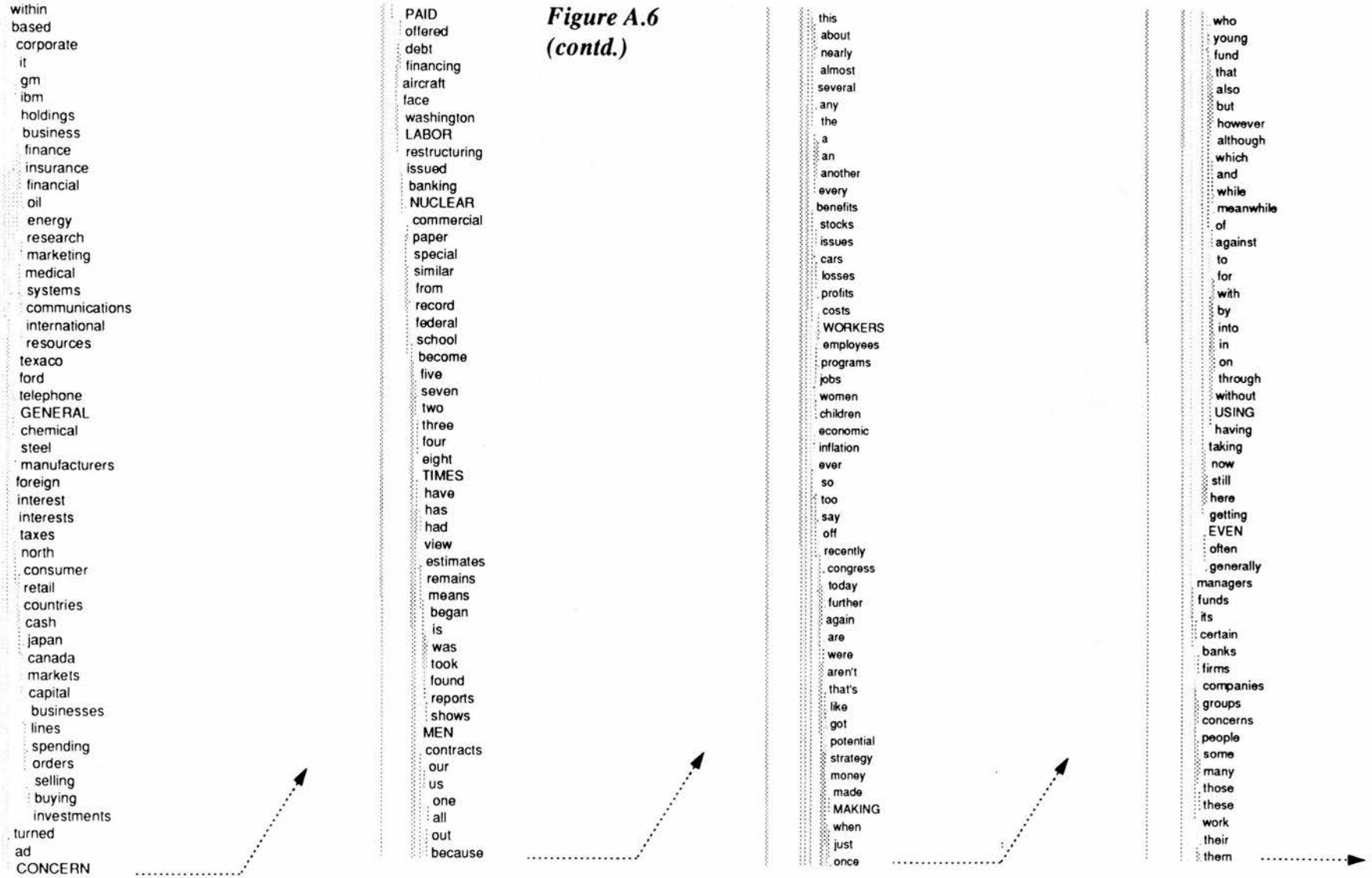


Figure A.6 (contd.)

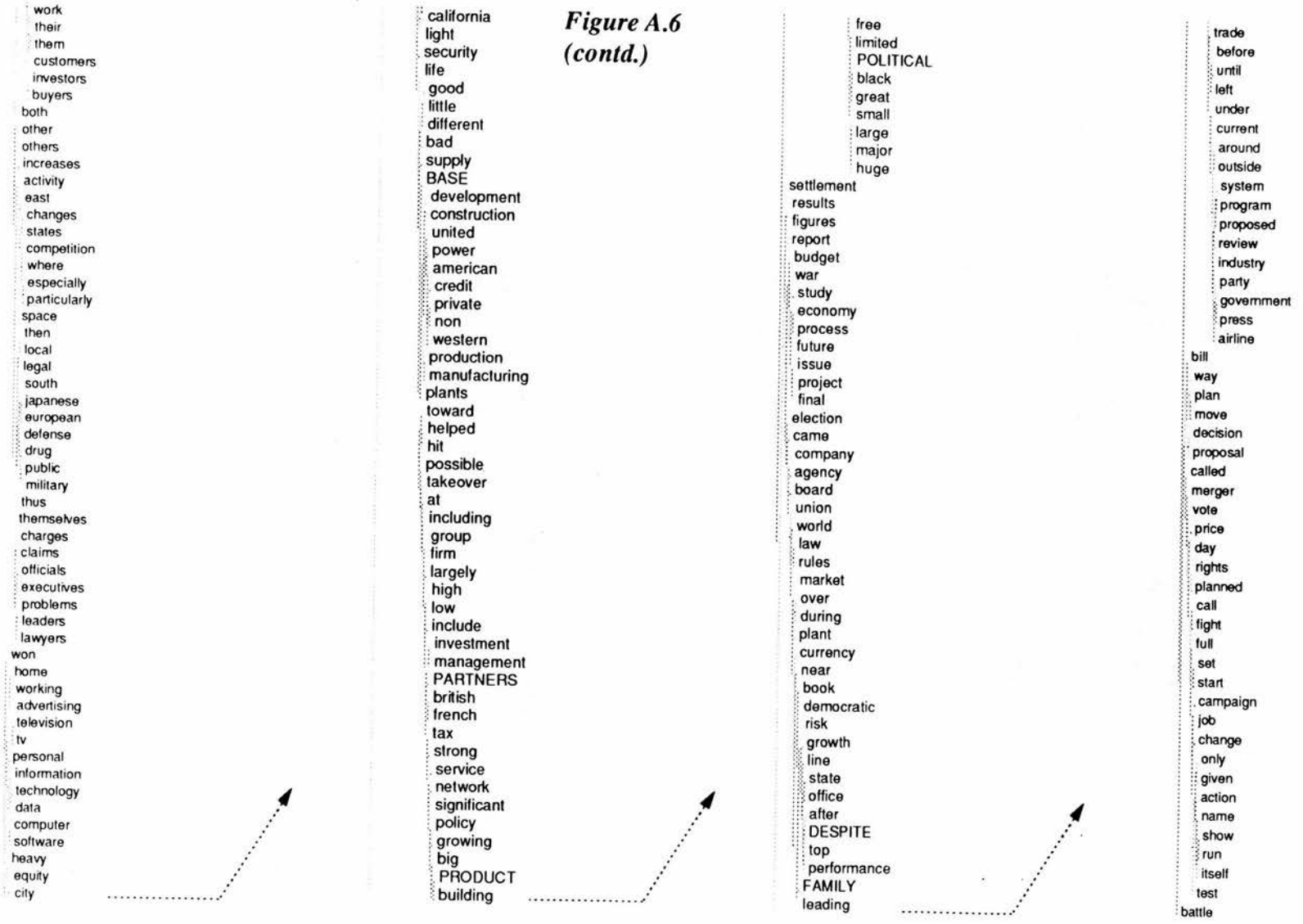


Figure A.6 (contd.)

Table A.7
(Spearman Distance Metric, Window Length=10)

The table below contains the 50 target words and 10 nearest neighbours for analysis 7 in Chapter 4.

Target Word	10 Nearest Neighbours (Spearman Correlation Coefficient)
able	can (0.800) they (0.794) if (0.792) even (0.786) that (0.783) could (0.783) have (0.782) so (0.782) but (0.778) any (0.777)
above	below (0.815) level (0.780) higher (0.768) average (0.764) low (0.761) price (0.759) than (0.757) points (0.754) year (0.751) from (0.747)
analyst	analysts (0.748) big (0.716) at (0.698) much (0.692) added (0.684) market (0.683) up (0.673) good (0.667) but (0.665) still (0.665)
base	well (0.698) and (0.697) more (0.695) cost (0.684) growing (0.684) line (0.683) as (0.681) most (0.677) the (0.674) some (0.672)
close	at (0.794) friday (0.780) yesterday (0.772) up (0.770) after (0.769) down (0.768) the (0.761) early (0.760) closed (0.759) trading (0.758)
concern	company (0.853) based (0.840) group (0.823) unit (0.814) holding (0.807) said (0.804) inc (0.801) its (0.794) previously (0.793) corp (0.788)
deal	out (0.768) be (0.761) any (0.756) he (0.752) mr (0.752) it (0.749) if (0.747) that (0.746) make (0.742) isn't (0.742)
despite	while (0.796) recent (0.782) strong (0.781) in (0.779) continued (0.778) analysts (0.777) as (0.776) the (0.775) however (0.773) on (0.769)
even	but (0.914) so (0.913) though (0.891) still (0.881) they (0.879) have (0.878) some (0.877) more (0.876) many (0.875) now (0.872)
expects	quarter (0.790) expected (0.788) net (0.775) earnings (0.772) third (0.765) its (0.758) fourth (0.758) profit (0.753) said (0.752) year (0.750)
family	his (0.709) who (0.708) mr (0.705) life (0.701) a (0.699) one (0.697) own (0.691) and (0.691) it (0.684) he (0.681)
gained	rose (0.788) fell (0.784) dropped (0.758) closed (0.740) volume (0.733) trading (0.726) share (0.725) ended (0.721) unchanged (0.713) sharply (0.712)
george	who (0.747) his (0.747) president (0.732) richard (0.727) john (0.727) robert (0.724) mr (0.724) old (0.722) former (0.716) david (0.715)
germany	west (0.766) german (0.727) japan (0.724) europe (0.704) european (0.694) japanese (0.673) world (0.667) foreign (0.648) government (0.647) east (0.646)
general	of (0.776) and (0.772) a (0.768) new (0.757) co (0.756) said (0.751) group (0.751) for (0.750) by (0.745) also (0.742)
hard	like (0.837) so (0.833) what (0.824) do (0.823) it's (0.822) too (0.820) way (0.817) not (0.811) people (0.806) good (0.803)
included	million (0.742) including (0.733) related (0.727) include (0.724) from (0.723) includes (0.719) which (0.717) by (0.707) of (0.705) also (0.702)
independent	it (0.680) own (0.678) has (0.677) members (0.674) public (0.671) mr (0.669) decision (0.666) outside (0.664) an (0.663) management (0.662)
index	stocks (0.801) prices (0.792) volume (0.789) traders (0.788) trading (0.769) points (0.768) decline (0.766) average (0.766) market (0.752) price (0.751)
it's	think (0.903) you (0.901) going (0.901) says (0.899) do (0.894) what (0.893) like (0.892) we (0.889) not (0.887) i (0.885)
labor	department (0.683) workers (0.680) washington (0.674) state (0.654) force (0.652) the (0.644) and (0.643) government (0.643) economic (0.636) health (0.635)
making	make (0.848) more (0.805) made (0.805) such (0.804) that (0.799) have (0.798) some (0.796) other (0.795) as (0.793) all (0.792)
men	women (0.780) people (0.765) who (0.761) him (0.747) his (0.744) she (0.740) her (0.735) man (0.732) i (0.731) white (0.728)
night	his (0.729) him (0.701) mr (0.701) out (0.693) he (0.692) then (0.685) when (0.685) white (0.683) time (0.681) here (0.680)
nuclear	power (0.698) military (0.686) defense (0.680) plant (0.677) project (0.658) soviet (0.651) force (0.648) plants (0.646) state (0.640) government (0.636)
old	who (0.829) his (0.790) mr (0.771) president (0.771) says (0.763) young (0.755) he (0.754) him (0.748) man (0.747) left (0.746)
operating	operations (0.820) profit (0.769) net (0.768) earnings (0.764) company (0.764) sales (0.759) company's (0.758) quarter (0.752) revenue (0.749) expects (0.743)

Table A.7 (contd.)

paid	pay (0.806) or (0.759) for (0.757) about (0.755) of (0.747) the (0.735) amount (0.735) had (0.725) a (0.724) each (0.723)
partners	group (0.738) firm (0.732) stake (0.708) management (0.706) company (0.706) investment (0.699) acquisition (0.690) shareholders (0.684) holding (0.683) investor (0.682)
percentage	average (0.785) rate (0.766) year (0.755) compared (0.735) points (0.729) higher (0.728) above (0.724) slightly (0.718) price (0.715) previous (0.715)
political	democratic (0.836) party (0.798) policy (0.788) bush (0.778) reagan (0.775) leaders (0.773) campaign (0.765) war (0.760) what (0.756) not (0.754)
preferred	holding (0.714) common (0.807) holders (0.795) outstanding (0.782) shares (0.769) transaction (0.744) purchase (0.741) offer (0.723) offering (0.720) company (0.714)
product	line (0.736) more (0.722) products (0.719) industry (0.714) computer (0.711) makers (0.705) growing (0.703) computers (0.702) costs (0.700) wide (0.700)
products	equipment (0.805) business (0.782) technology (0.780) parts (0.764) systems (0.763) businesses (0.757) maker (0.754) industries (0.754) manufacturing (0.752) computer (0.751)
same	only (0.852) more (0.835) than (0.830) all (0.822) that (0.822) but (0.817) have (0.815) even (0.810) one (0.808) the (0.804)
should	not (0.867) if (0.860) can (0.858) that (0.850) could (0.84C) have (0.845) they (0.844) would (0.843) be (0.832) even (0.832)
take	that (0.845) if (0.838) have (0.833) go (0.833) make (0.832) but (0.829) out (0.828) they (0.826) be (0.825) say (0.820)
they're	it's (0.843) you (0.831) don't (0.831) think (0.824) going (0.824) like (0.823) do (0.821) says (0.816) so (0.809) people (0.808)
times	time (0.745) than (0.742) one (0.739) most (0.732) as (0.729) only (0.729) just (0.727) more (0.726) the (0.725) all (0.725)
transaction	acquisition (0.828) company (0.798) purchase (0.795) offer (0.794) sale (0.793) merger (0.767) stake (0.767) shareholders (0.761) holders (0.754) holding (0.750)
transportation	department (0.676) energy (0.673) services (0.667) general (0.650) service (0.645) international (0.639) air (0.638) new (0.636) said (0.635) also (0.632)
use	such (0.820) used (0.806) using (0.792) make (0.771) can (0.768) provide (0.764) other (0.762) help (0.759) have (0.758) their (0.755)
using	use (0.792) used (0.782) such (0.772) can (0.745) into (0.741) have (0.736) more (0.736) their (0.735) other (0.733) own (0.729)
wall	street (0.985) journal (0.909) news (0.773) york (0.730) new (0.717) wsj (0.712) staff (0.710) offers (0.704) stk (0.700) international (0.700)
wednesday	friday (0.810) monday (0.794) yesterday (0.787) tuesday (0.783) late (0.754) week (0.752) trading (0.728) exchange (0.720) closed (0.712) month (0.701)
week	month (0.836) last (0.827) friday (0.806) yesterday (0.797) late (0.784) after (0.780) monday (0.775) weeks (0.773) three (0.771) the (0.769)
weren't	disclosed (0.737) terms (0.725) said (0.724) concern (0.721) based (0.721) company (0.721) business (0.718) which (0.709) held (0.706) also (0.703)
will	would (0.871) to (0.845) may (0.840) be (0.834) next (0.827) could (0.818) won't (0.816) and (0.813) the (0.811) of (0.809)
william	john (0.807) robert (0.784) david (0.781) james (0.776) richard (0.761) president (0.758) chairman (0.754) chief (0.749) executive (0.748) jr (0.747)
workers	employees (0.769) jobs (0.760) work (0.758) plant (0.712) force (0.702) cost (0.700) those (0.697) plants (0.691) benefits (0.691) only (0.690)

Figure A.7 below shows the dendrogram containing the 1000 target words considered in analysis 7 of Chapter 4.

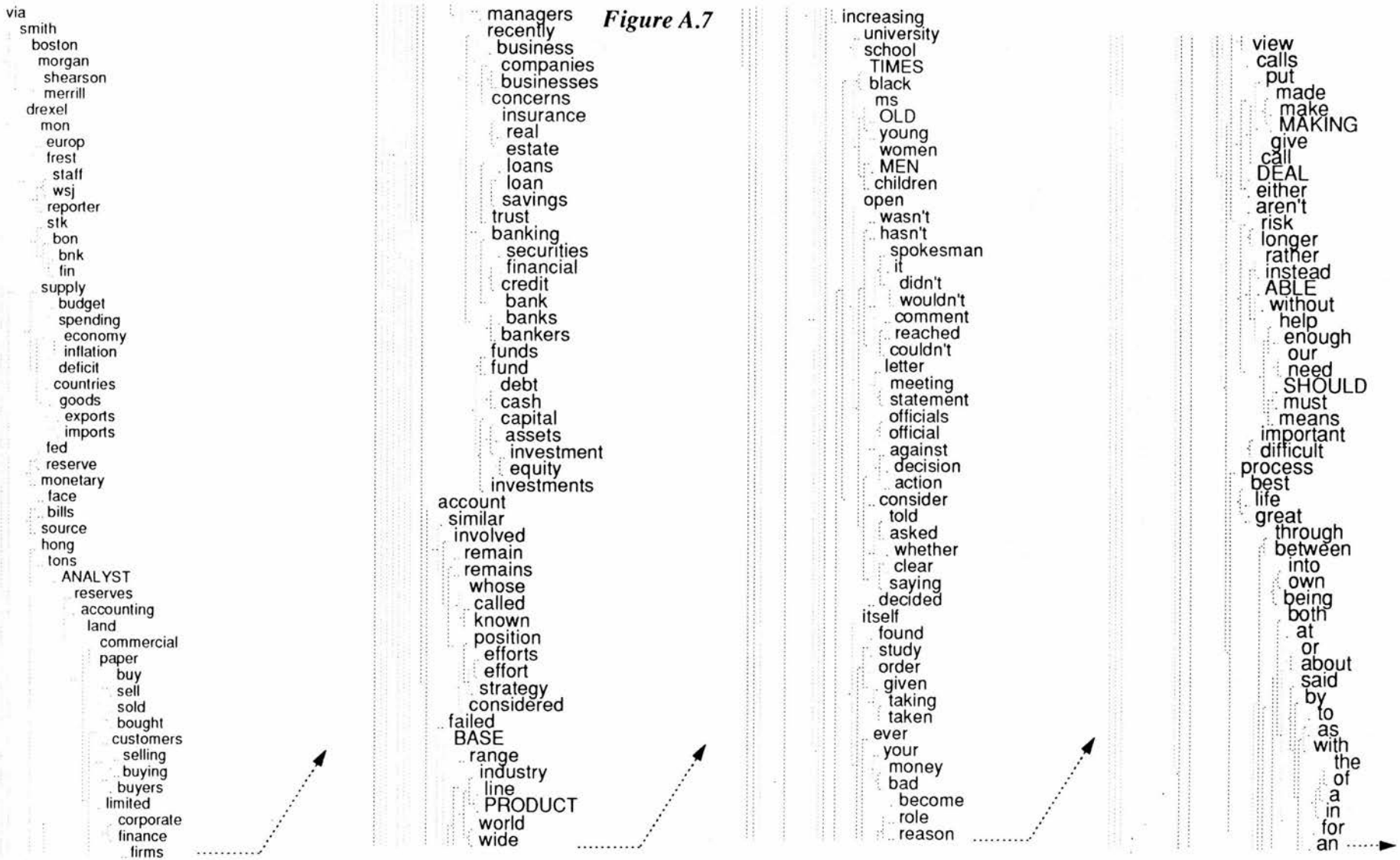

via
smith
boston
morgan
shearson
merrill
drexel
mon
europ
frest
staff
wsj
reporter
stk
bon
bnk
fin
supply
budget
spending
economy
inflation
deficit
countries
goods
exports
imports
fed
reserve
monetary
face
bills
source
hong
tons
ANALYST
reserves
accounting
land
commercial
paper
buy
sell
sold
bought
customers
selling
buying
buyers
limited
corporate
finance
firms
managers
recently
business
companies
businesses
concerns
insurance
real
estate
loans
loan
savings
trust
banking
securities
financial
credit
bank
banks
bankers
funds
fund
debt
cash
capital
assets
investment
equity
investments
account
similar
involved
remain
remains
whose
called
known
position
efforts
effort
strategy
considered
failed
BASE
range
industry
line
PRODUCT
world
wide
increasing
university
school
TIMES
black
ms
OLD
young
women
MEN
children
open
wasn't
hasn't
spokesman
it
didn't
wouldn't
comment
reached
couldn't
letter
meeting
statement
officials
official
against
decision
action
consider
told
asked
whether
clear
saying
decided
itself
found
study
order
given
taking
taken
ever
your
money
bad
become
role
reason
view
calls
put
made
make
MAKING
give
call
DEAL
either
aren't
risk
longer
rather
instead
ABLE
without
help
enough
our
need
SHOULD
must
means
important
difficult
process
best
life
great
through
between
into
own
being
both
at
or
about
said
by
to
as
with
the
of
a
in
for
an

Figure A.7

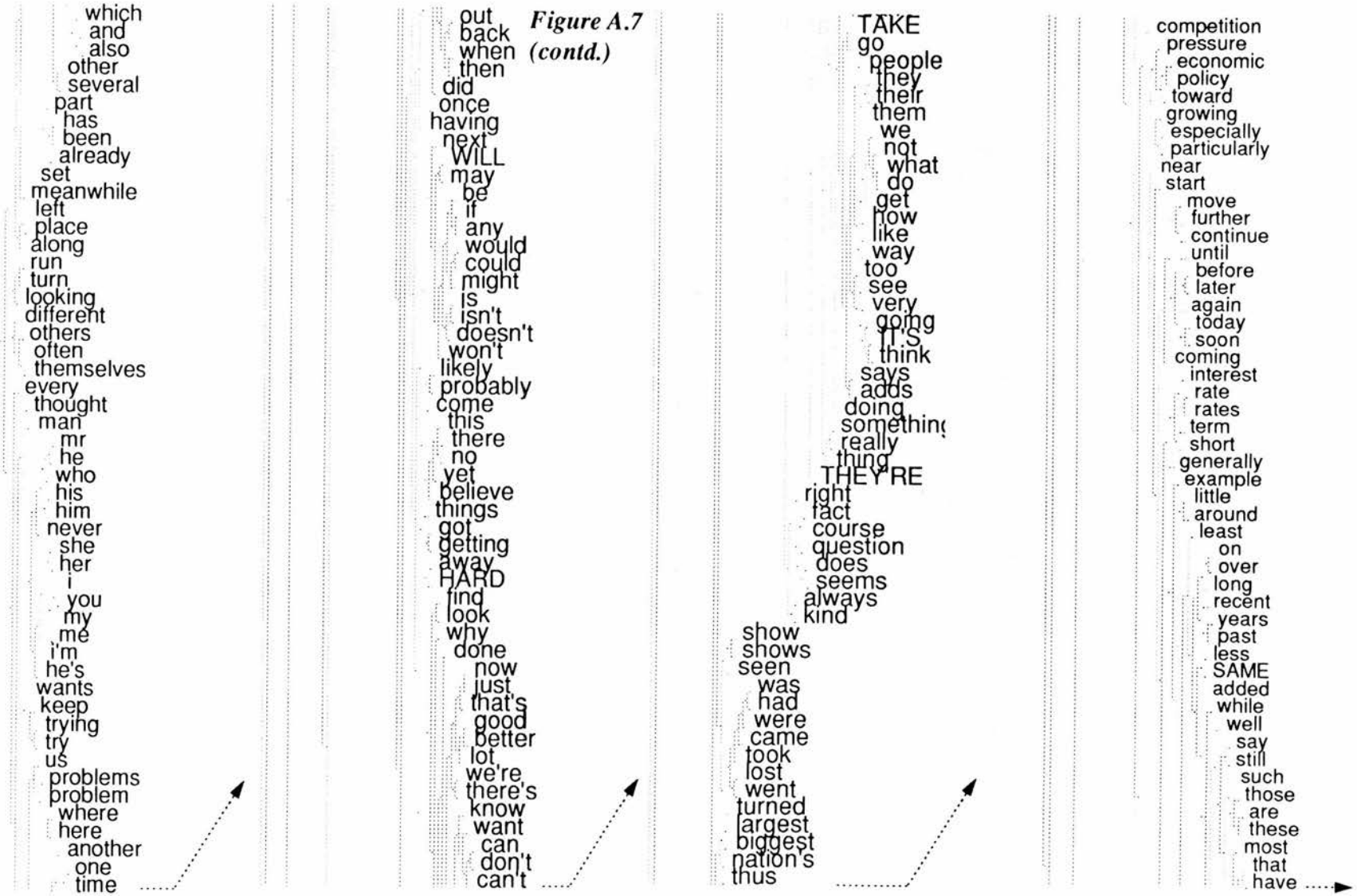


Figure A.7 (contd.)

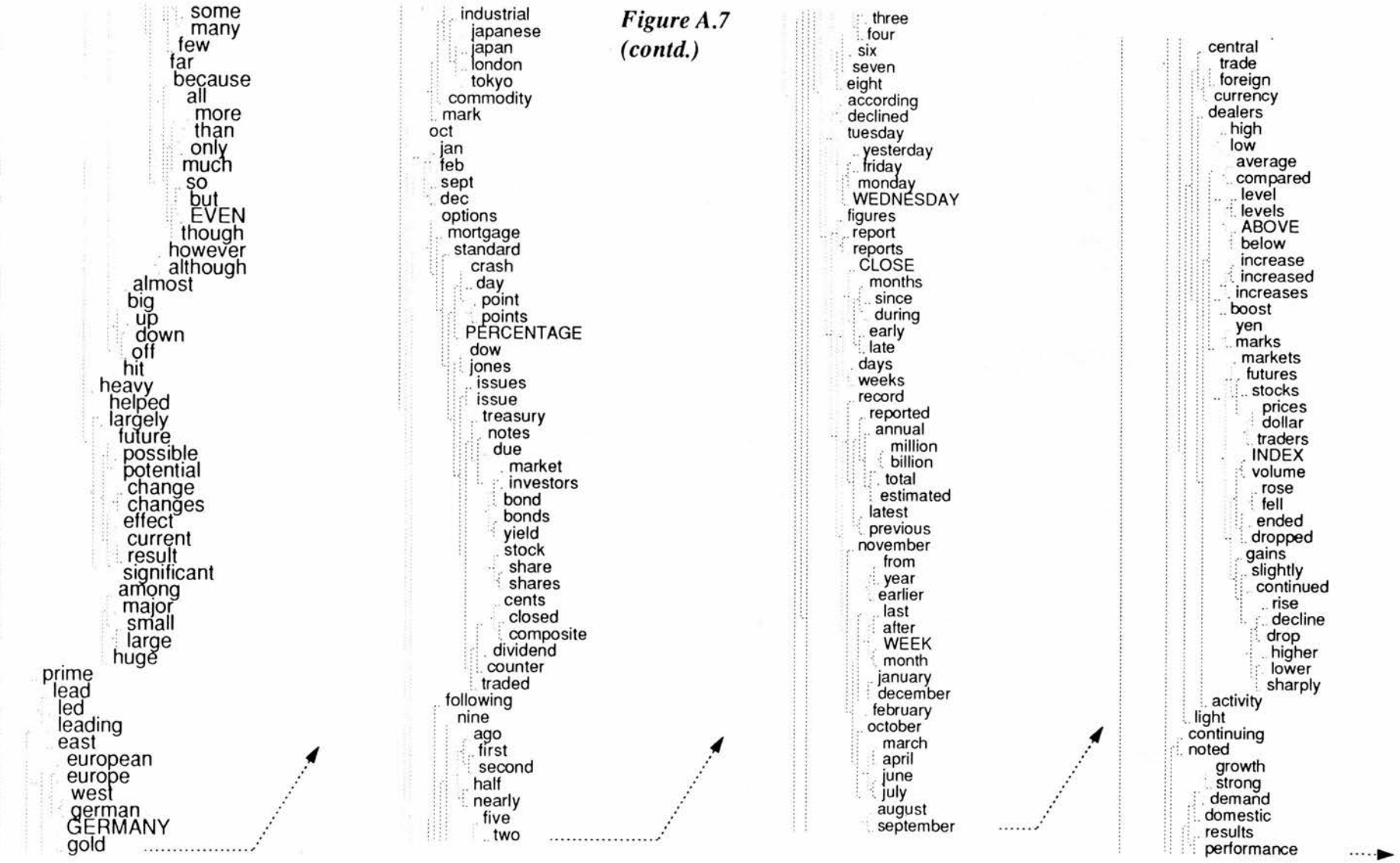


Figure A.7 (contd.)

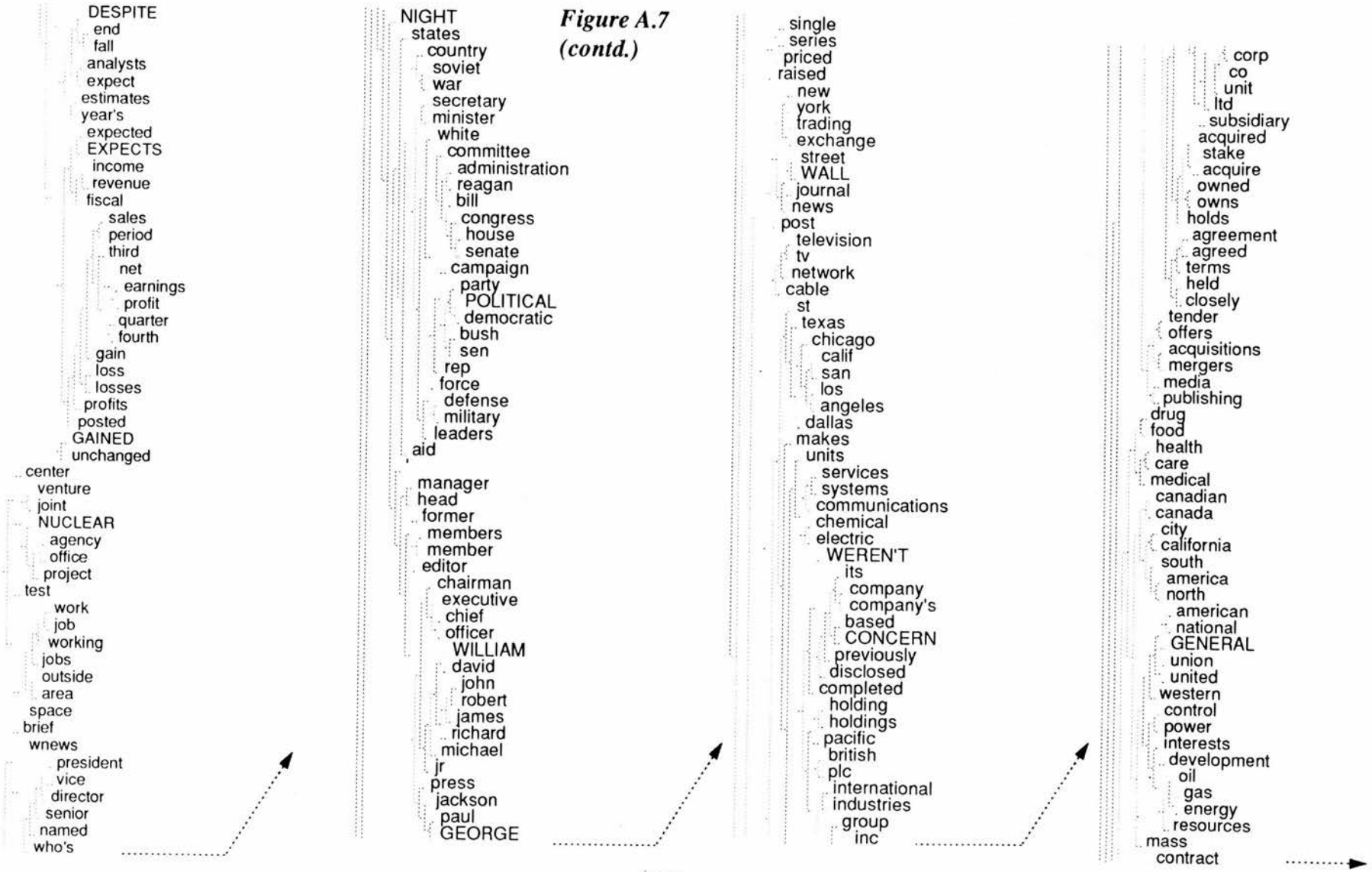


Figure A.7 (contd.)

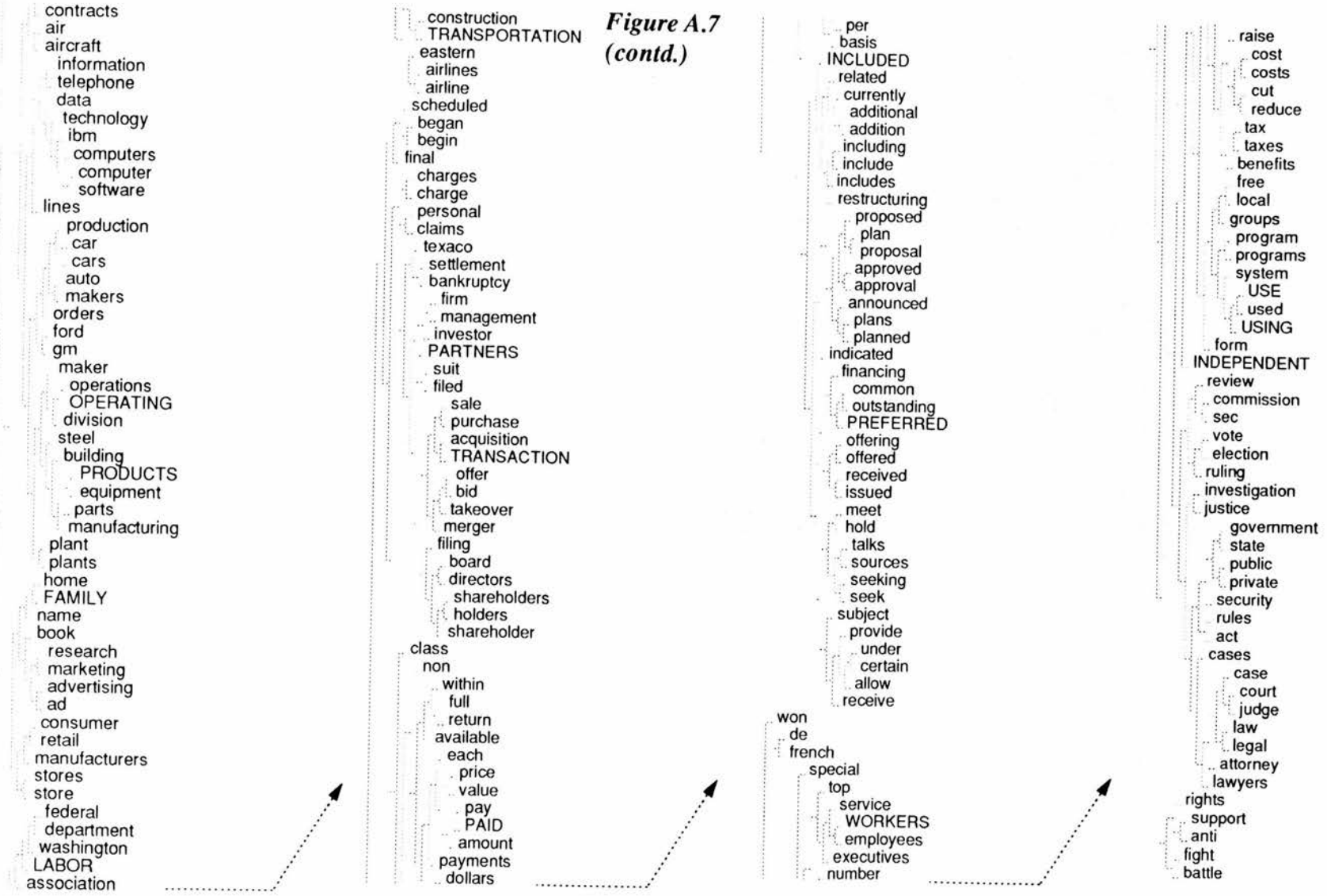


Figure A.7 (contd.)

Table A.8
(Euclidean Distance Metric, Window Length=10)

The table below contains the 50 target words and 10 nearest neighbours for analysis 8 in Chapter 4.

Target Word	10 Nearest Neighbours (Euclidean Distance)
able	try (0.032) allow (0.032) going (0.033) trying (0.033) meet (0.033) continue (0.033) enough (0.034) likely (0.034) seems (0.034) wants (0.034)
above	below (0.023) issue (0.024) during (0.025) level (0.025) rate (0.026) price (0.027) final (0.027) over (0.028) before (0.028) current (0.028)
analyst	smith (0.039) manager (0.042) investor (0.042) official (0.043) drexel (0.043) san (0.043) mass (0.044) co (0.044) investment (0.044) merrill (0.044)
base	line (0.019) huge (0.020) high (0.020) product (0.020) in (0.020) large (0.020) public (0.020) through (0.020) top (0.020) also (0.020)
close	dropped (0.029) up (0.030) down (0.031) back (0.031) move (0.031) open (0.032) start (0.032) lost (0.032) plan (0.032) begin (0.032)
concern	group (0.018) management (0.021) firm (0.021) for (0.021) building (0.021) energy (0.021) with (0.022) investment (0.022) merger (0.022) including (0.022)
deal	way (0.019) show (0.019) is (0.020) problem (0.020) without (0.020) move (0.020) line (0.020) having (0.021) run (0.021) strategy (0.021)
despite	in (0.012) on (0.013) of (0.014) performance (0.014) after (0.014) however (0.014) industry (0.014) although (0.015) against (0.015) by (0.015)
even	but (0.012) now (0.013) still (0.013) that (0.013) just (0.013) so (0.014) once (0.014) to (0.014) without (0.014) not (0.015)
expects	plans (0.031) agreed (0.034) sell (0.035) seek (0.036) raise (0.037) boost (0.038) acquire (0.038) dropped (0.038) wants (0.038) expect (0.039)
family	public (0.016) state (0.017) book (0.017) office (0.017) leading (0.018) great (0.018) political (0.019) huge (0.019) through (0.019) top (0.019)
gained	dropped (0.035) according (0.036) fell (0.036) sell (0.036) try (0.037) agreed (0.038) trying (0.038) close (0.039) plans (0.039) allow (0.039
george	richard (0.017) james (0.018) robert (0.018) john (0.018) michael (0.019) william (0.019) paul (0.019) david (0.020) former (0.021) marketing (0.023)
germany	german (0.031) japan (0.039) canada (0.040) west (0.040) east (0.041) south (0.042) europe (0.043) western (0.043) investments (0.043) home (0.043)
general	financial (0.018) american (0.018) business (0.018) united (0.018) data (0.018) energy (0.019) management (0.019) international (0.019) service (0.019) a (0.019)
hard	get (0.017) find (0.017) make (0.017) them (0.018) work (0.018) go (0.018) enough (0.019) see (0.019) how (0.019) help (0.019)
included	gain (0.022) charge (0.025) year (0.027) million (0.028) or (0.029) profit (0.029) period (0.030) reported (0.030) from (0.030) continuing (0.031)
independent	outside (0.015) also (0.017) potential (0.018) all (0.018) public (0.018) through (0.019) the (0.019) certain (0.019) out (0.019) anti (0.019)
index	points (0.033) futures (0.037) off (0.039) market (0.039) monday (0.040) dollar (0.040) tuesday (0.041) stocks (0.041) price (0.041) issues (0.041)
it's	like (0.015) that's (0.015) good (0.016) there's (0.017) very (0.017) just (0.017) we're (0.018) something (0.019) little (0.019) see (0.019)
labor	by (0.019) washington (0.020) and (0.020) defense (0.021) workers (0.022) in (0.022) to (0.022) the (0.022) with (0.022) for (0.022)
making	with (0.011) also (0.012) now (0.012) to (0.012) that (0.013) having (0.013) made (0.013) using (0.013) and (0.013) for (0.013)
men	women (0.020) groups (0.022) children (0.022) leaders (0.022) political (0.023) left (0.023) those (0.024) are (0.024) people (0.024) all (0.024)
night	was (0.021) after (0.021) when (0.021) made (0.021) in (0.022) took (0.022) with (0.022) hit (0.022) on (0.022) around (0.022)
nuclear	power (0.018) plant (0.020) public (0.021) by (0.022) military (0.022) which (0.023) state (0.023) system (0.023) of (0.023) review (0.023)
old	who (0.026) this (0.026) five (0.026) former (0.028) last (0.028) young (0.028) with (0.028) two (0.028) job (0.029) four (0.029)
operating	operations (0.025) financial (0.029) including (0.030) chairman (0.030) finance (0.031) production (0.031) services (0.031) products (0.032) post (0.032) energy (0.032)

Table A.8 (contd.)

paid	about (0.019) also (0.019) made (0.019) which (0.020) for (0.020) full (0.020) given (0.020) set (0.020) in (0.020) two (0.020)
partners	management (0.016) investment (0.016) limited (0.017) a (0.019) recently (0.019) an (0.019) firm (0.020) its (0.020) american (0.020) with (0.020)
percentage	rate (0.034) above (0.034) point (0.036) seven (0.036) average (0.036) full (0.036) half (0.037) level (0.037) change (0.037) leading (0.037)
political	public (0.014) policy (0.015) left (0.015) military (0.015) leaders (0.016) into (0.016) great (0.016) outside (0.016) is (0.016) only (0.016)
preferred	common (0.028) holders (0.037) shares (0.038) offering (0.039) class (0.039) purchase (0.039) each (0.039) cash (0.040) receive (0.041) dividend (0.042)
product	line (0.014) making (0.015) strategy (0.016) building (0.016) with (0.016) also (0.016) potential (0.016) huge (0.016) in (0.016) test (0.016)
products	equipment (0.017) businesses (0.018) services (0.018) food (0.018) marketing (0.019) technology (0.019) lines (0.019) medical (0.019) systems (0.020) manufacturing (0.020)
same	economy (0.024) during (0.026) company's (0.027) biggest (0.027) process (0.028) government (0.028) nation's (0.028) final (0.029) level (0.029) act (0.030)
should	must (0.010) could (0.013) may (0.014) will (0.014) would (0.014) might (0.016) won't (0.016) can (0.017) however (0.021) not (0.021)
take	make (0.011) help (0.013) give (0.013) get (0.014) go (0.015) keep (0.016) lead (0.016) find (0.016) be (0.016) hold (0.017)
they're	aren't (0.018) they (0.018) see (0.019) like (0.019) people (0.019) that's (0.020) it's (0.020) just (0.020) often (0.020) we're (0.020)
times	in (0.018) on (0.019) of (0.019) two (0.020) around (0.020) top (0.020) big (0.020) almost (0.020) for (0.020) to (0.020)
transaction	company (0.018) proposed (0.022) project (0.023) final (0.023) company's (0.024) proposal (0.025) merger (0.025) settlement (0.025) agreement (0.026) issue (0.026)
transportation	off (0.028) energy (0.030) financial (0.031) technology (0.032) defense (0.032) labor (0.032) service (0.032) data (0.032) non (0.033) industrial (0.033)
use	support (0.016) help (0.016) work (0.016) certain (0.017) hold (0.017) lead (0.017) used (0.017) aid (0.018) rights (0.018) them (0.019)
using	making (0.013) having (0.013) free (0.013) also (0.014) run (0.014) to (0.014) without (0.015) and (0.015) work (0.015) meanwhile (0.015)
wall	journal (0.062) street (0.066) staff (0.068) europ (0.070) offers (0.072) frest (0.072) publishing (0.079) news (0.079) stk (0.080) acquisitions (0.080)
wednesday	tuesday (0.013) monday (0.016) friday (0.024) late (0.024) up (0.024) down (0.024) heavy (0.026) slightly (0.026) april (0.027) march (0.027)
week	month (0.018) fall (0.023) night (0.024) after (0.025) in (0.026) for (0.027) to (0.027) later (0.027) on (0.027) april (0.027)
weren't	also (0.029) and (0.029) officials (0.029) to (0.030) in (0.030) were (0.030) the (0.030) data (0.030) with (0.030) several (0.030)
will	may (0.014) should (0.014) would (0.014) must (0.014) could (0.015) won't (0.017) to (0.017) however (0.018) also (0.018) through (0.018)
william	richard (0.008) robert (0.009) john (0.009) james (0.010) david (0.012) michael (0.014) paul (0.014) jr (0.015) george (0.019) former (0.021)
workers	employees (0.013) jobs (0.015) companies (0.016) local (0.017) costs (0.017) are (0.017) other (0.017) all (0.017) work (0.018) groups (0.018)

Figure A.8 below shows the dendrogram containing the 1000 target words considered in analysis 8 of Chapter 4.

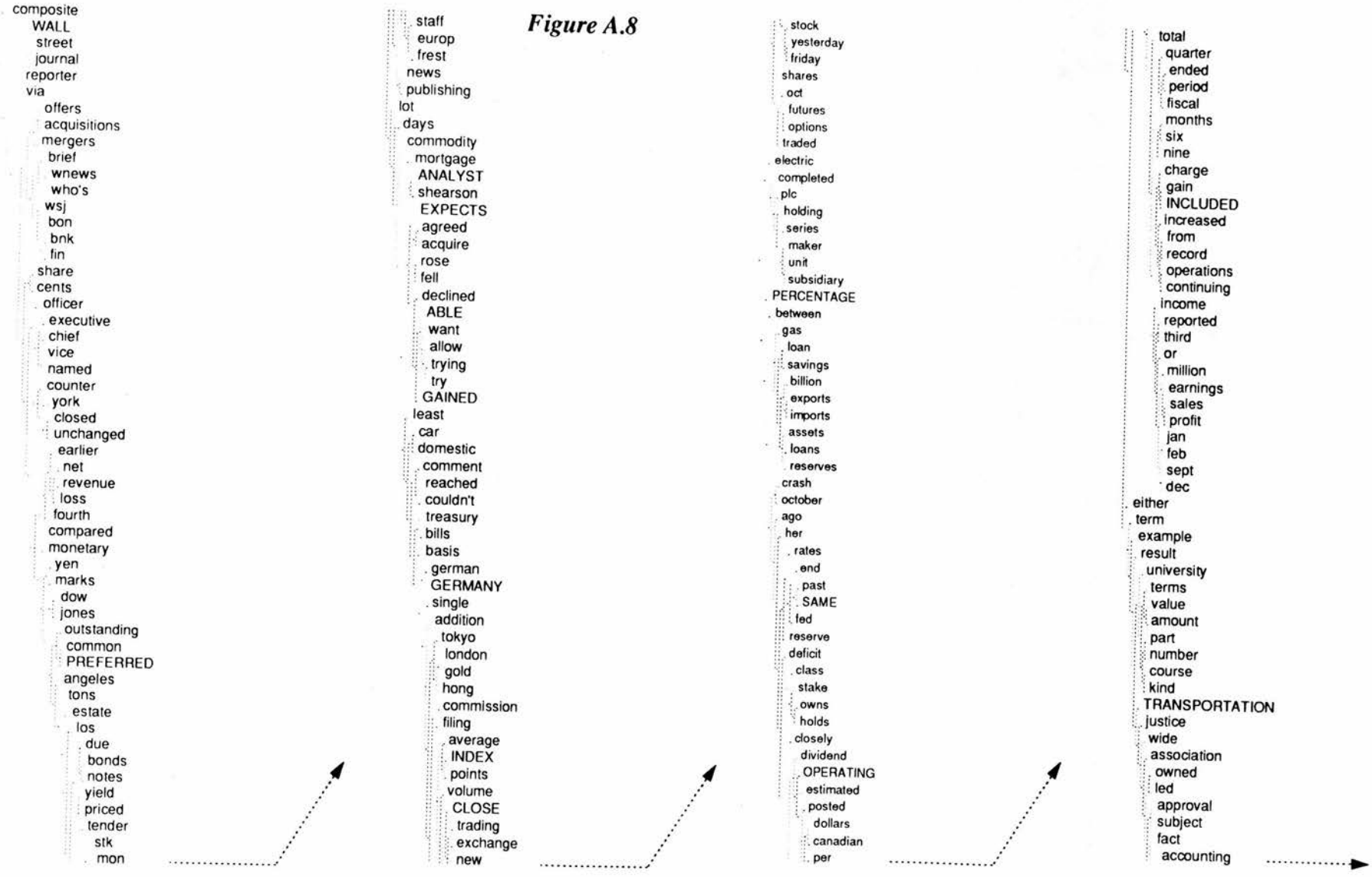

composite
WALL
street
journal
reporter
via
offers
acquisitions
mergers
brief
wnews
who's
wsj
bon
bnk
fin
share
cents
officer
executive
chief
vice
named
counter
york
closed
unchanged
earlier
net
revenue
loss
fourth
compared
monetary
yen
marks
dow
jones
outstanding
common
PREFERRED
angeles
tons
estate
los
due
bonds
notes
yield
priced
tender
stk
mon
staff
europ
frest
news
publishing
lot
days
commodity
mortgage
ANALYST
shearson
EXPECTS
agreed
acquire
rose
fell
declined
ABLE
want
allow
trying
try
GAINED
least
car
domestic
comment
reached
couldn't
treasury
bills
basis
german
GERMANY
single
addition
tokyo
london
gold
hong
commission
filing
average
INDEX
points
volume
CLOSE
trading
exchange
new
stock
yesterday
friday
shares
oct
futures
options
traded
electric
completed
plc
holding
series
maker
unit
subsidiary
PERCENTAGE
between
gas
loan
savings
billion
exports
imports
assets
loans
reserves
crash
october
ago
her
rates
end
past
SAME
fed
reserve
deficit
class
stake
owns
holds
closely
dividend
OPERATING
estimated
posted
dollars
canadian
per
total
quarter
ended
period
fiscal
months
six
nine
charge
gain
INCLUDED
increased
from
record
operations
continuing
income
reported
third
or
million
earnings
sales
profit
jan
feb
sept
dec
either
term
example
result
university
terms
value
amount
part
number
course
kind
TRANSPORTATION
justice
wide
association
owned
led
approval
subject
fact
accounting

Figure A.8

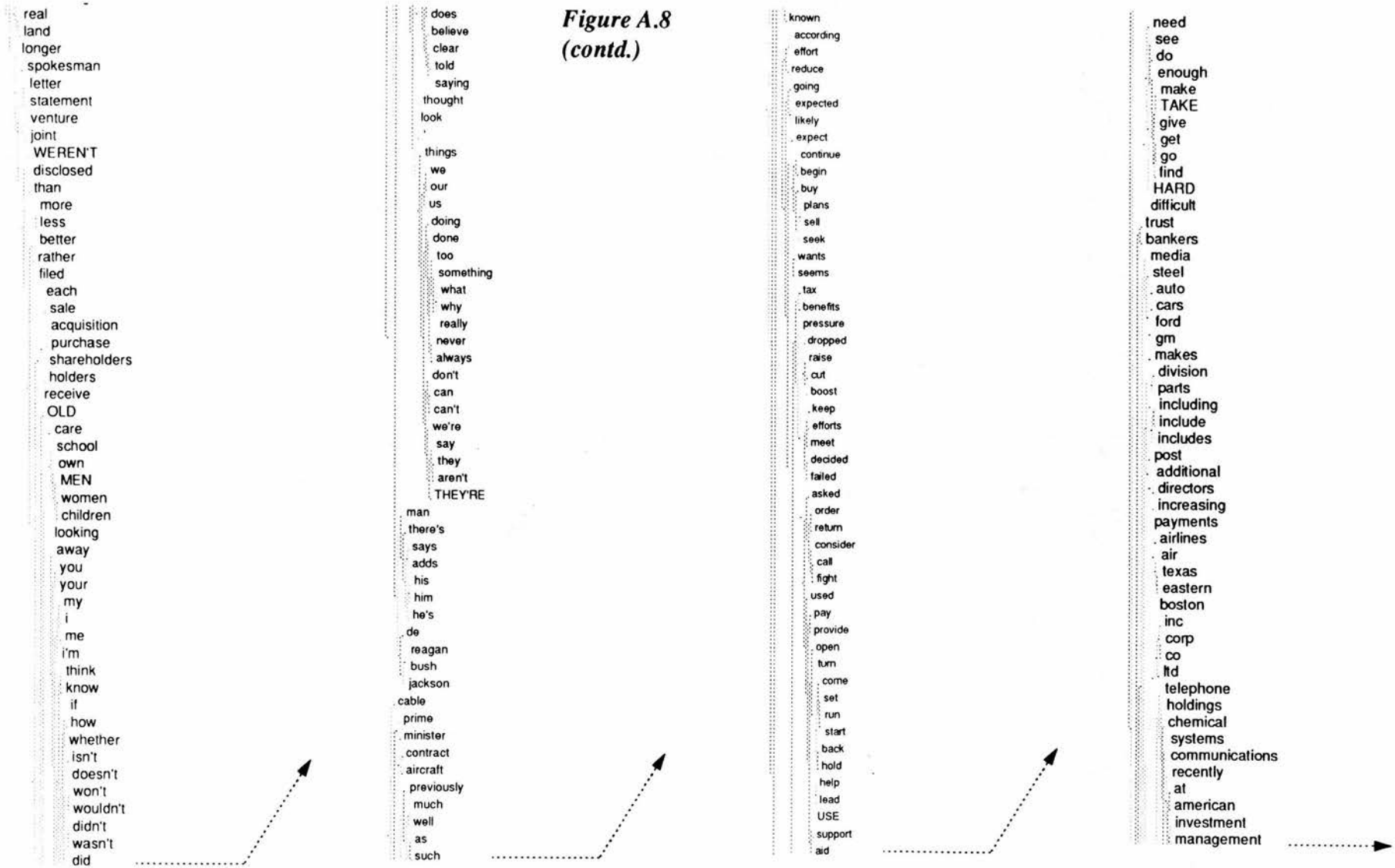


Figure A.8 (contd.)

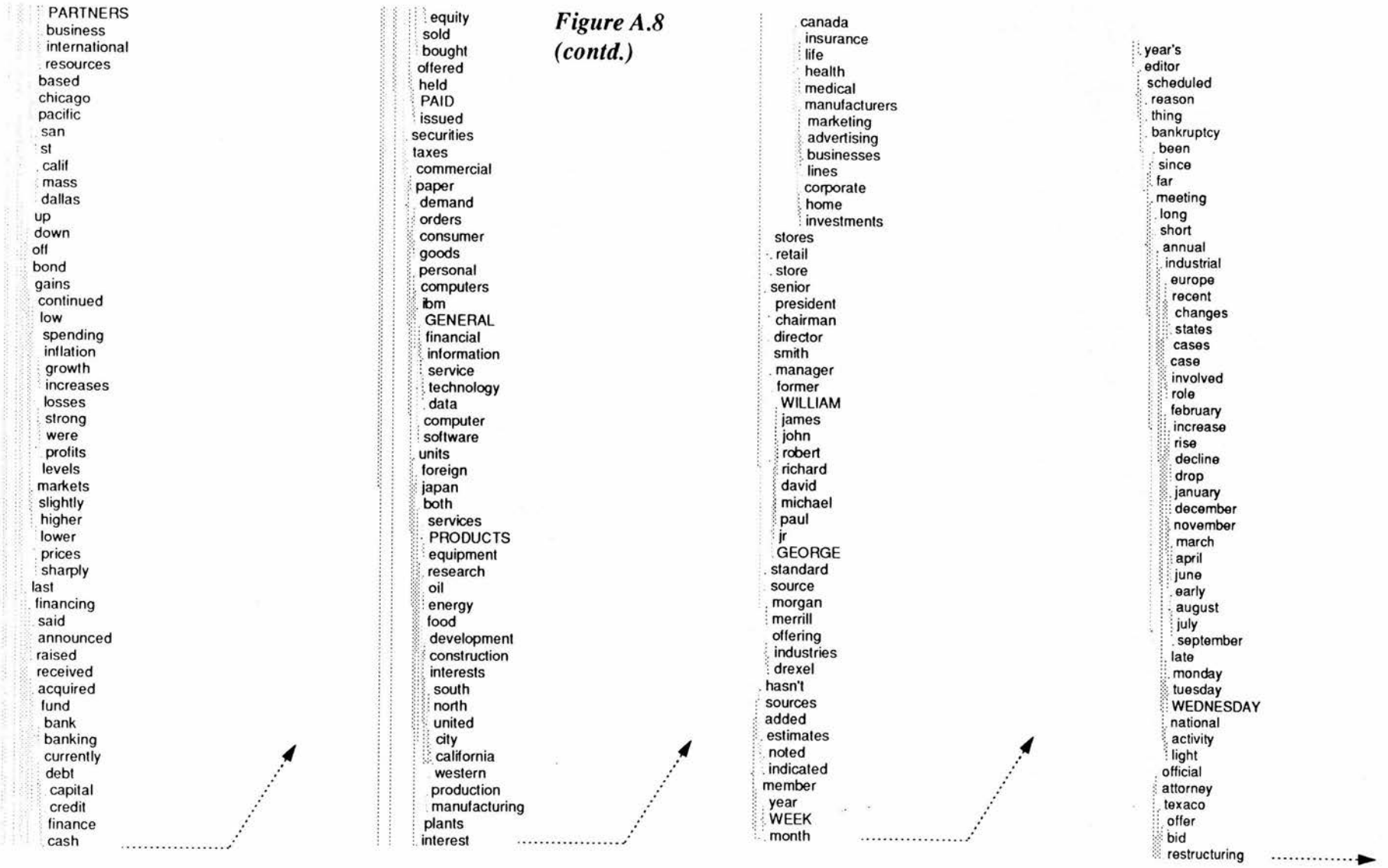


Figure A.8 (contd.)

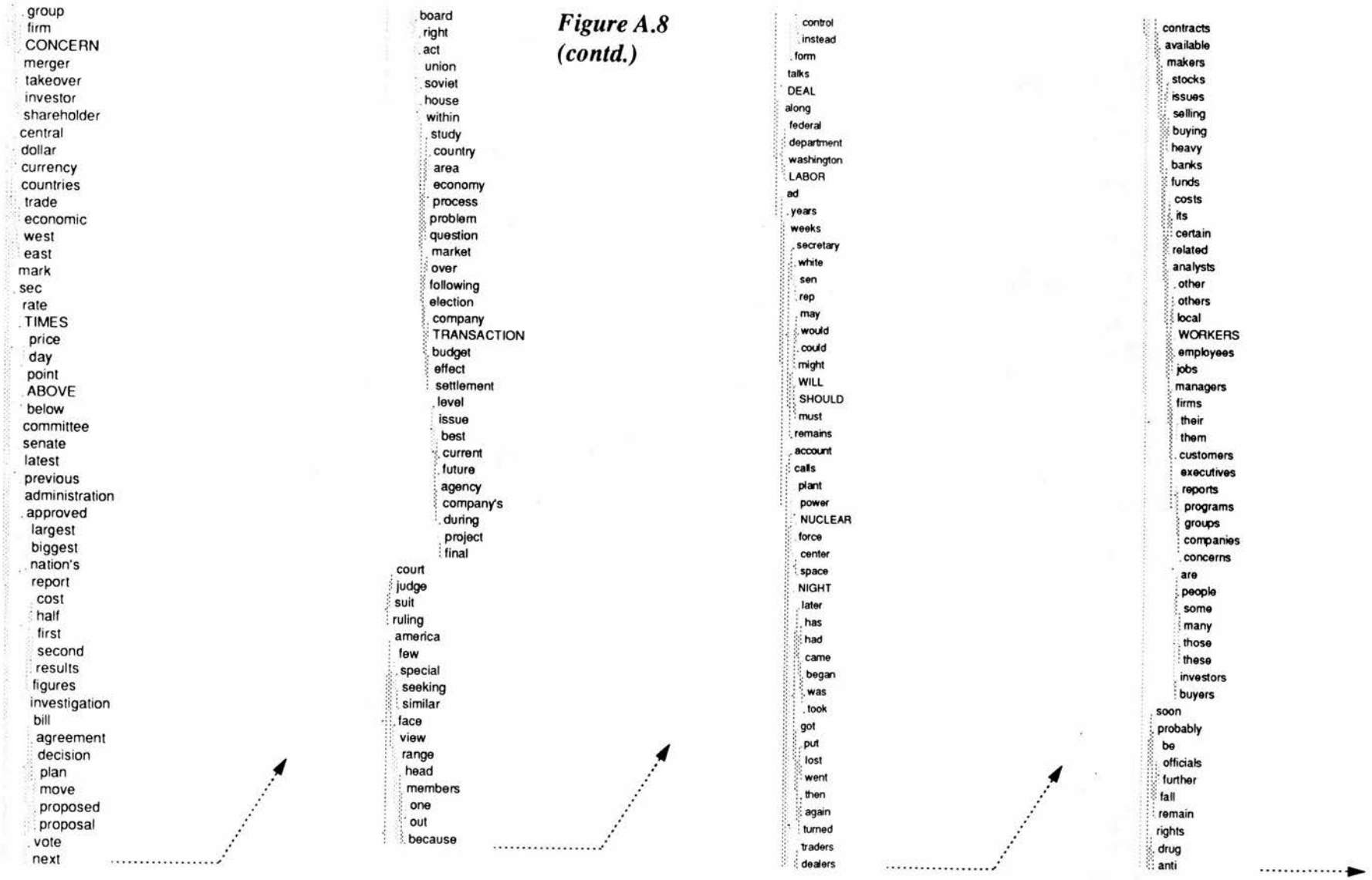


Figure A.8 (contd.)

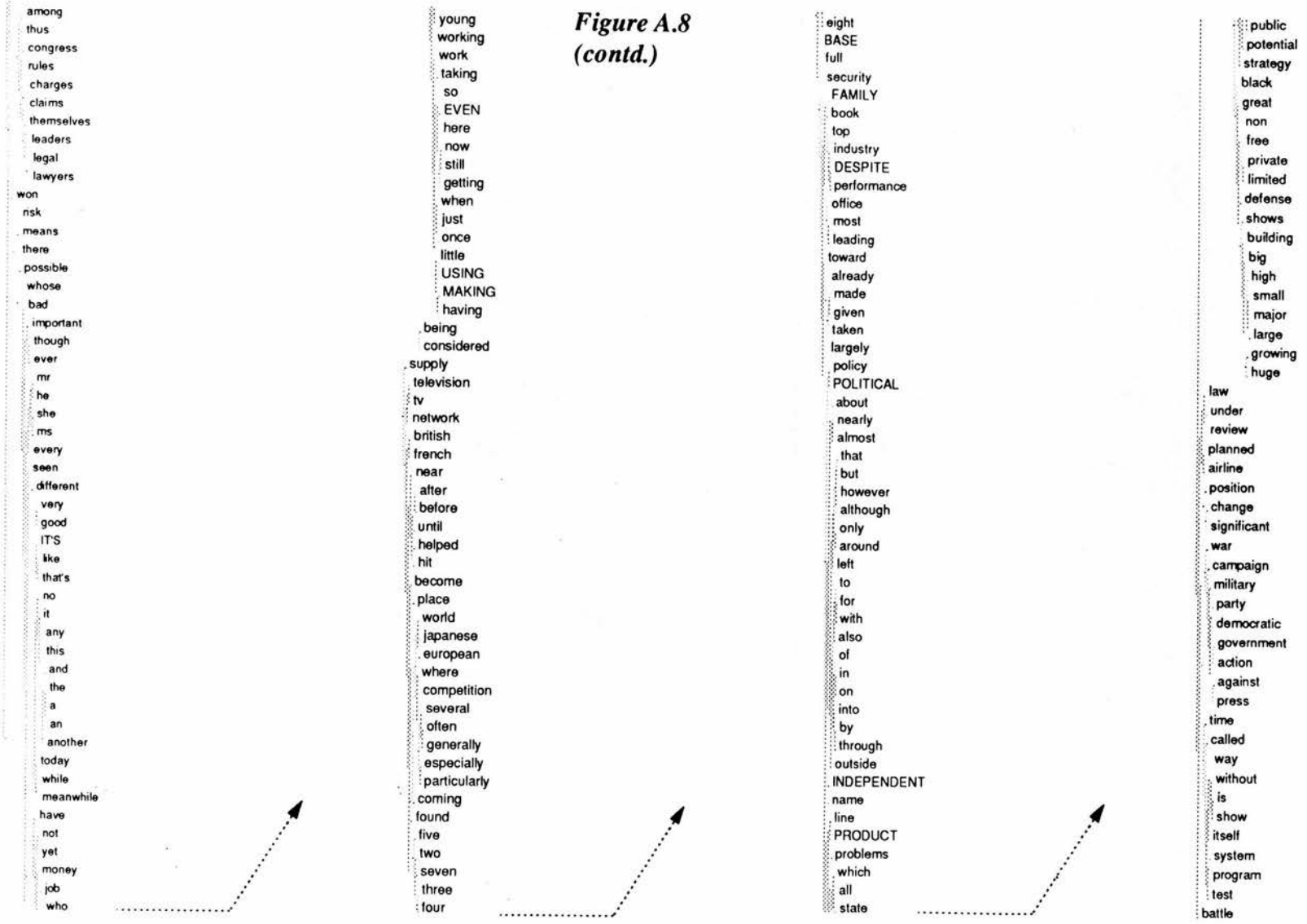


Figure A.8 (contd.)

Table A.9
(Spearman Distance Metric, Window Length=25)

The table below contains the 50 target words and 10 nearest neighbours for analysis 9 in Chapter 4.

Target Word	10 Nearest Neighbours (Spearman Correlation Coefficient)
able	have (0.881) if (0.876) could (0.873) make (0.873) that (0.873) can (0.872) be (0.869) they (0.869) but (0.868) such (0.868)
above	below (0.854) points (0.848) higher (0.846) point (0.841) level (0.837) slightly (0.830) average (0.830) at (0.826) low (0.826) down (0.823)
analyst	analysts (0.858) big (0.804) added (0.792) expect (0.790) because (0.787) market (0.782) much (0.781) performance (0.780) up (0.779) buying (0.777)
base	more (0.786) well (0.784) past (0.779) as (0.777) large (0.776) than (0.775) low (0.774) most (0.773) and (0.771) high (0.770)
close	friday (0.858) yesterday (0.855) at (0.854) closed (0.844) trading (0.842) up (0.839) down (0.836) monday (0.835) after (0.833) stock (0.832)
concern	based (0.905) company (0.904) group (0.892) inc (0.890) unit (0.889) previously (0.887) its (0.885) holding (0.881) said (0.880) corp (0.879)
deal	out (0.847) any (0.844) be (0.843) if (0.837) isn't (0.836) might (0.830) it (0.824) is (0.824) that (0.823) being (0.822)
despite	while (0.881) strong (0.878) recent (0.872) in (0.867) fall (0.864) continued (0.863) level (0.863) up (0.862) during (0.860) than (0.860)
even	so (0.955) but (0.949) though (0.942) have (0.938) now (0.938) only (0.935) still (0.935) they (0.933) many (0.930) some (0.928)
expects	quarter (0.874) net (0.867) earnings (0.864) said (0.858) fourth (0.854) expected (0.853) third (0.851) profit (0.850) cents (0.849) its (0.848)
family	own (0.794) a (0.792) whose (0.790) it (0.787) has (0.786) who (0.784) mr (0.783) life (0.781) one (0.781) and (0.780)
gained	fell (0.844) rose (0.839) dropped (0.835) volume (0.820) closed (0.813) trading (0.801) unchanged (0.795) sharply (0.792) gain (0.791) share (0.784)
george	who (0.854) he (0.848) his (0.842) mr (0.834) i (0.816) old (0.816) him (0.809) michael (0.807) president (0.807) this (0.807)
germany	west (0.830) german (0.812) japan (0.802) europe (0.798) european (0.795) world (0.769) foreign (0.760) japanese (0.747) government (0.741) central (0.736)
general	of (0.844) and (0.839) a (0.836) for (0.830) new (0.828) two (0.825) by (0.824) co (0.824) also (0.816) recently (0.809)
hard	so (0.903) like (0.895) too (0.891) way (0.889) what (0.889) do (0.888) even (0.885) says (0.885) it's (0.885) just (0.884)
included	million (0.836) from (0.835) related (0.831) which (0.828) including (0.823) reported (0.817) includes (0.813) also (0.811) earlier (0.811) last (0.809)
independent	has (0.812) it (0.801) decision (0.801) control (0.795) any (0.794) an (0.790) under (0.788) own (0.788) seeking (0.787) members (0.786)
index	stocks (0.864) volume (0.854) traders (0.837) points (0.836) prices (0.835) average (0.835) market (0.819) fell (0.818) decline (0.816) rise (0.810)
it's	says (0.948) think (0.943) going (0.943) do (0.941) like (0.941) what (0.938) you (0.938) so (0.932) we (0.930) there's (0.930)
labor	workers (0.784) department (0.765) the (0.750) force (0.748) washington (0.747) government (0.745) in (0.745) for (0.745) by (0.744) and (0.743)
making	make (0.905) that (0.891) is (0.884) have (0.882) has (0.882) but (0.881) with (0.880) more (0.880) other (0.880) made (0.879)
men	women (0.842) who (0.840) people (0.832) young (0.830) his (0.827) him (0.823) man (0.820) her (0.808) school (0.808) never (0.805)
night	his (0.816) him (0.801) out (0.795) never (0.789) then (0.788) he (0.784) place (0.781) who (0.780) i (0.779) white (0.779)
nuclear	power (0.772) military (0.761) defense (0.753) plant (0.732) soviet (0.721) force (0.720) project (0.718) state (0.713) government (0.704) administration (0.703)
old	who (0.895) his (0.853) mr (0.846) he (0.838) left (0.837) president (0.832) young (0.828) years (0.824) director (0.820) him (0.819)
operating	operations (0.887) company's (0.839) officer (0.838) net (0.837) expects (0.837) company (0.837) profit (0.828) loss (0.826) business (0.822) revenue (0.821)

Table A.9 (contd.)

paid	pay (0.871) or (0.834) for (0.833) of (0.821) a (0.817) about (0.812) payments (0.811) the (0.809) amount (0.807) had (0.806)
partners	group (0.812) firm (0.808) stake (0.797) company (0.796) held (0.790) owns (0.788) management (0.788) owned (0.785) agreed (0.782) shareholders (0.780)
percentage	average (0.843) rate (0.832) year (0.819) higher (0.804) above (0.793) compared (0.789) lower (0.787) below (0.782) slightly (0.782) points (0.777)
political	democratic (0.905) party (0.877) reagan (0.865) leaders (0.855) bush (0.855) policy (0.851) white (0.841) seems (0.834) sen (0.833) war (0.831)
preferred	common (0.883) holders (0.875) outstanding (0.869) shares (0.830) transaction (0.825) purchase (0.822) holding (0.808) acquisition (0.803) sale (0.796) cash (0.796)
product	line (0.823) research (0.816) industry (0.806) makers (0.797) computers (0.794) marketing (0.791) products (0.790) computer (0.789) more (0.789) wide (0.789)
products	business (0.855) equipment (0.850) maker (0.834) its (0.830) technology (0.822) businesses (0.821) manufacturing (0.819) company (0.818) industries (0.816) operations (0.816)
same	only (0.922) more (0.913) all (0.902) but (0.901) than (0.899) have (0.899) that (0.892) still (0.892) though (0.890) much (0.889)
should	not (0.919) if (0.916) have (0.915) that (0.913) can (0.909) could (0.898) they (0.898) so (0.896) even (0.895) now (0.895)
take	that (0.915) have (0.911) if (0.906) but (0.905) out (0.903) could (0.900) be (0.898) they (0.894) is (0.893) make (0.893)
they're	it's (0.901) says (0.897) don't (0.892) going (0.885) you (0.884) think (0.884) like (0.881) that's (0.881) do (0.881) lot (0.880)
times	time (0.831) most (0.829) one (0.820) than (0.820) when (0.815) but (0.814) as (0.813) more (0.813) just (0.812) around (0.812)
transaction	acquisition (0.906) acquire (0.879) purchase (0.877) stake (0.872) acquired (0.864) company (0.863) sale (0.857) offer (0.857) disclosed (0.855) holding (0.848)
transportation	energy (0.759) of (0.750) by (0.750) which (0.749) new (0.746) its (0.744) said (0.743) services (0.742) industries (0.740) including (0.739)
use	such (0.873) using (0.872) used (0.870) can (0.846) have (0.838) make (0.835) is (0.835) system (0.834) more (0.833) other (0.832)
using	use (0.872) such (0.848) used (0.841) can (0.841) into (0.826) have (0.825) example (0.824) now (0.823) their (0.823) are (0.822)
wall	street (0.996) journal (0.970) staff (0.882) reporter (0.877) news (0.865) new (0.817) wsj (0.812) york (0.811) international (0.794) offers (0.793)
wednesday	friday (0.875) monday (0.863) yesterday (0.854) tuesday (0.848) trading (0.827) late (0.821) exchange (0.816) week (0.804) closed (0.803) unchanged (0.785)
week	month (0.868) last (0.859) yesterday (0.841) friday (0.839) weeks (0.834) late (0.829) three (0.828) previous (0.826) at (0.825) april (0.825)
weren't	business (0.841) said (0.831) based (0.830) disclosed (0.828) company (0.819) held (0.816) group (0.815) concern (0.815) its (0.815) closely (0.814)
will	be (0.905) to (0.902) and (0.892) of (0.892) next (0.890) would (0.886) also (0.884) for (0.883) the (0.879) a (0.878)
william	john (0.882) robert (0.865) richard (0.855) david (0.854) james (0.850) president (0.837) paul (0.835) michael (0.833) director (0.830) jr (0.830)
workers	work (0.824) jobs (0.820) employees (0.806) plant (0.791) labor (0.784) plants (0.775) force (0.773) benefits (0.771) only (0.765) cost (0.764)

Figure A.9 below shows the dendrogram containing the 1000 target words considered in analysis 9 of Chapter 4.

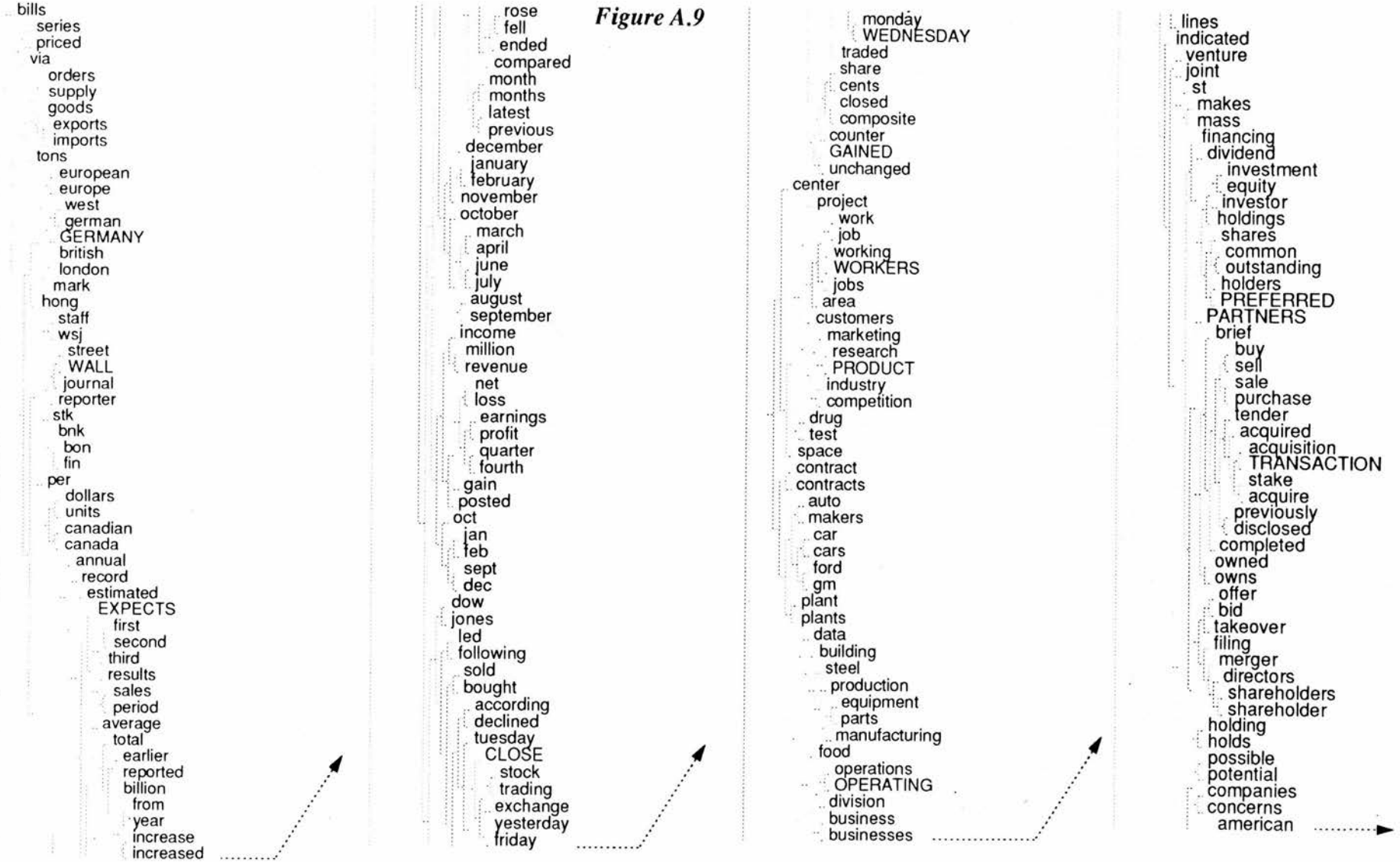

bills
series
priced
via
orders
supply
goods
exports
imports
tons
european
europe
west
german
GERMANY
british
london
mark
hong
staff
wsj
street
WALL
journal
reporter
stk
bnk
bon
fin
per
dollars
units
canadian
canada
annual
record
estimated
EXPECTS
first
second
third
results
sales
period
average
total
earlier
reported
billion
from
year
increase
increased
rose
fell
ended
compared
month
months
latest
previous
december
january
february
november
october
march
april
june
july
august
september
income
million
revenue
net
loss
earnings
profit
quarter
fourth
gain
posted
oct
jan
feb
sept
dec
dow
jones
led
following
sold
bought
according
declined
tuesday
CLOSE
stock
trading
exchange
yesterday
friday
monday
WEDNESDAY
traded
share
cents
closed
composite
counter
GAINED
unchanged
center
project
work
job
working
WORKERS
jobs
area
customers
marketing
research
PRODUCT
industry
competition
drug
test
space
contract
contracts
auto
makers
car
cars
ford
gm
plant
plants
data
building
steel
production
equipment
parts
manufacturing
food
operations
OPERATING
division
business
businesses
lines
indicated
venture
joint
st
makes
mass
financing
dividend
investment
equity
investor
holdings
shares
common
outstanding
holders
PREFERRED
PARTNERS
brief
buy
sell
sale
purchase
tender
acquired
acquisition
TRANSACTION
stake
acquire
previously
disclosed
completed
owned
owns
offer
bid
takeover
filing
merger
directors
shareholders
shareholder
holding
holds
possible
potential
companies
concerns
american

Figure A.9

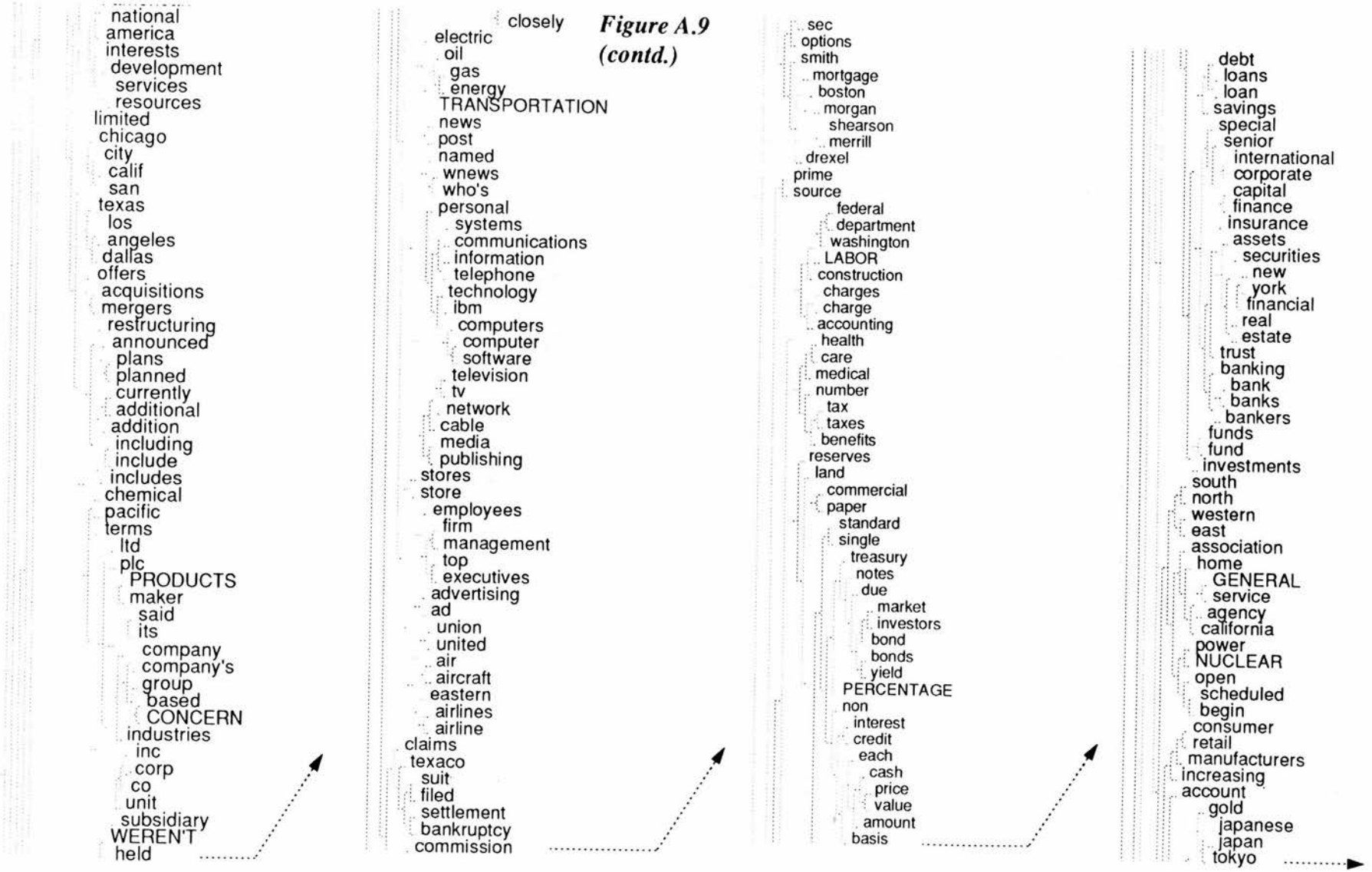


Figure A.9 (contd.)

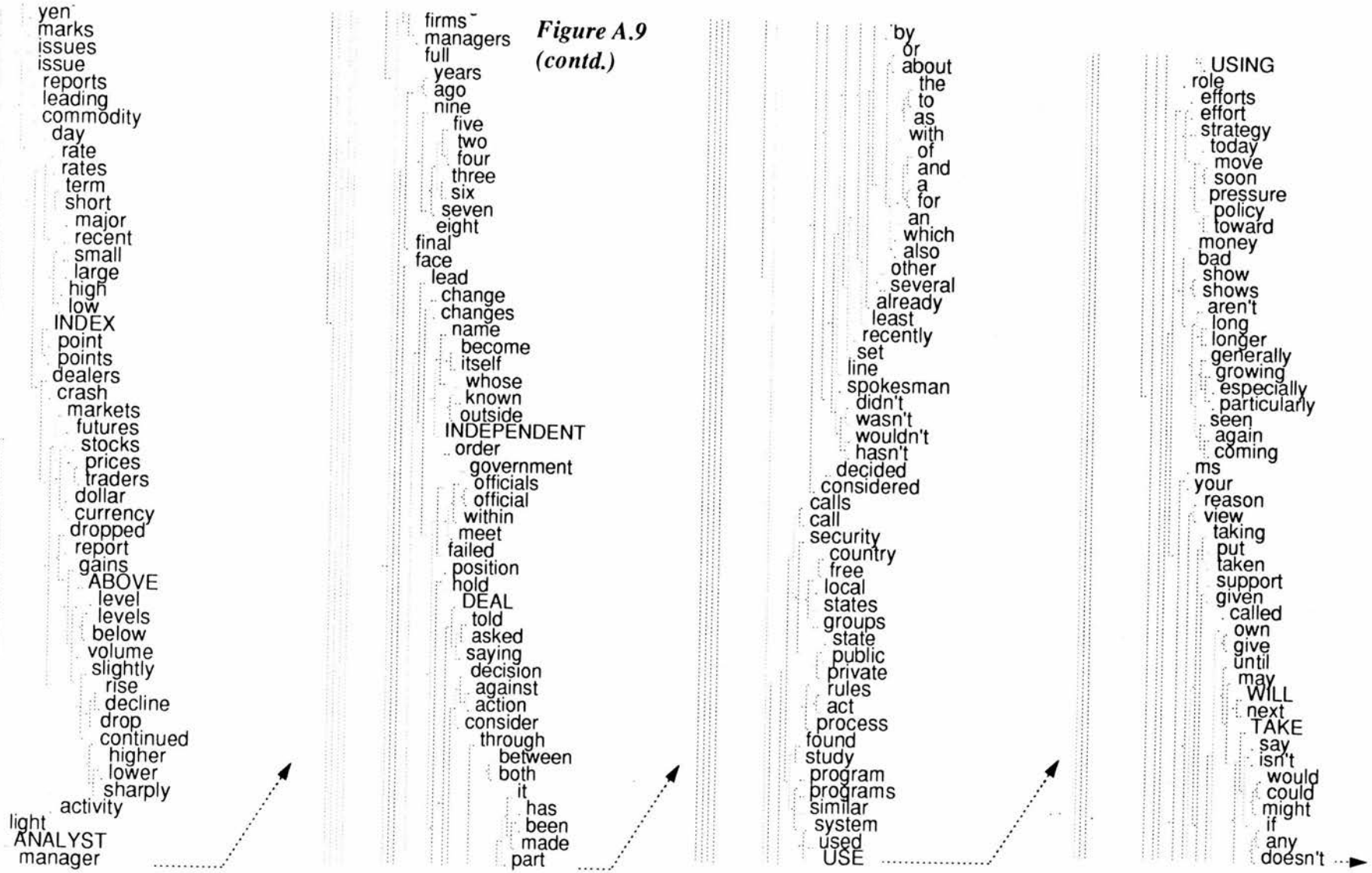


Figure A.9 (contd.)

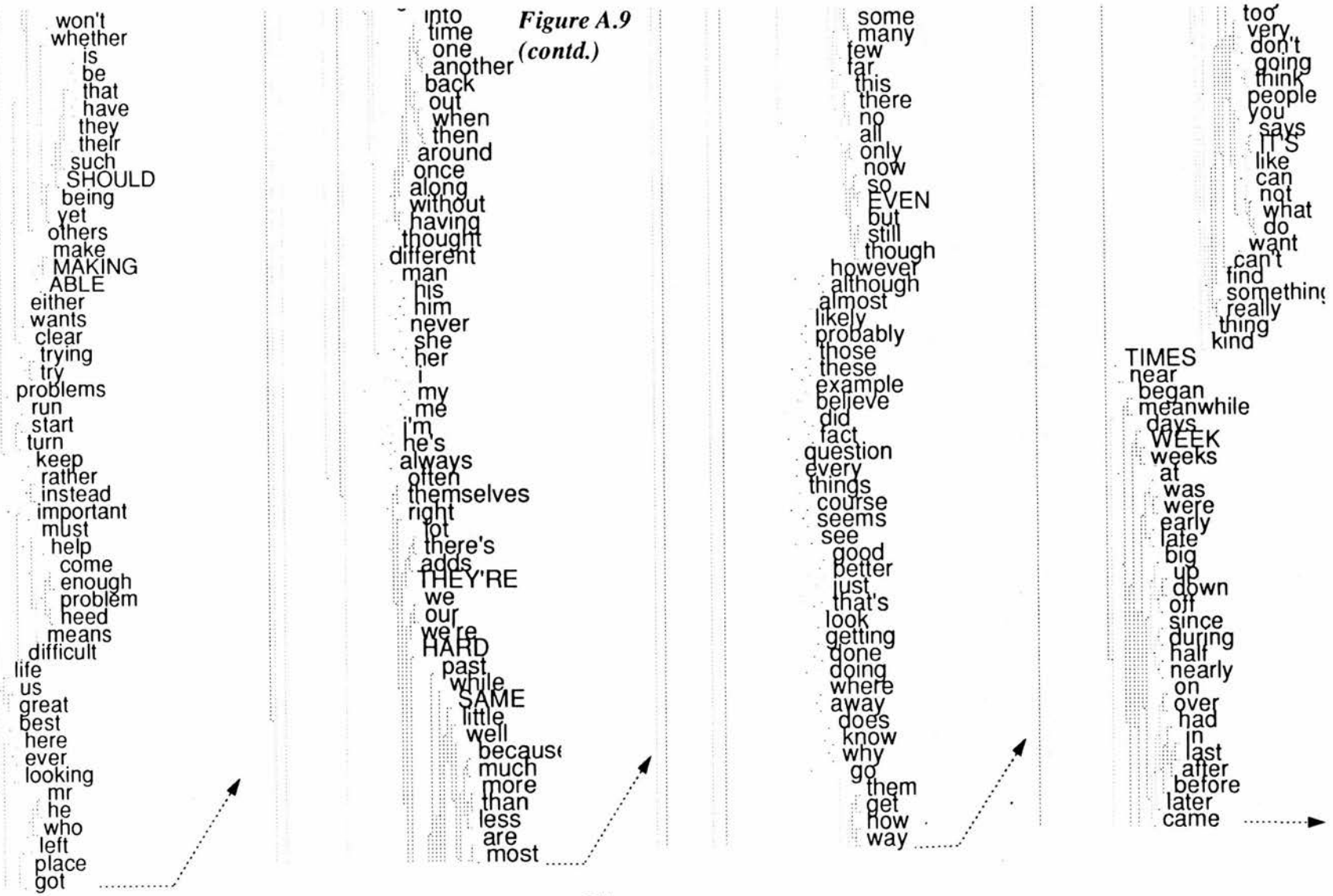


Figure A.9 (contd.)

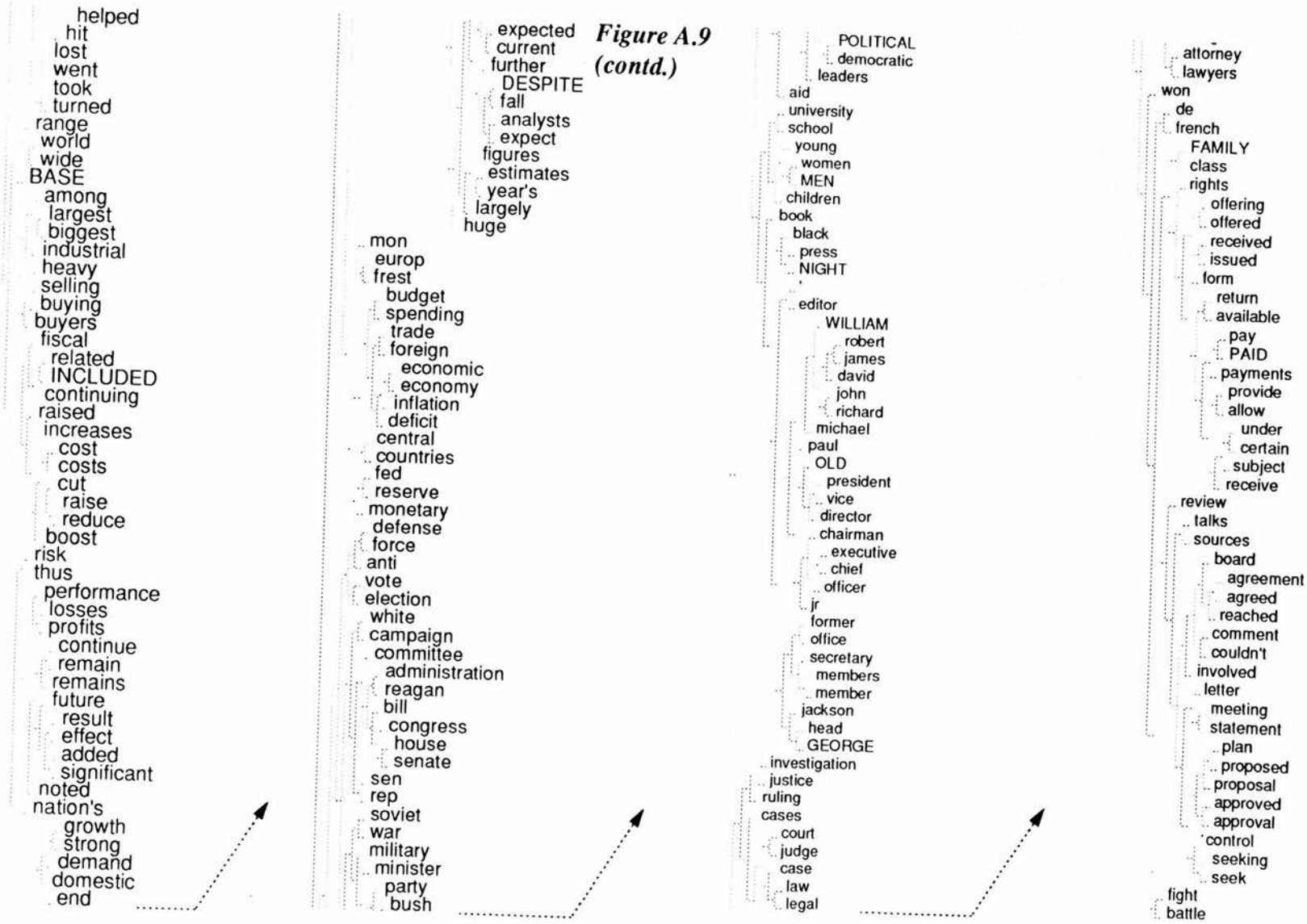


Figure A.9 (contd.)

Table A.10
(Euclidean Distance Metric, Window Length=25)

The table below contains the 50 target words and 10 nearest neighbours for analysis 10 in Chapter 4.

Target Word	10 Nearest Neighbours (Euclidean Distance)
able	try (0.015) enough (0.015) likely (0.016) meet (0.016) trying (0.016) allow (0.016) keep (0.016) take (0.016) wants (0.016) going (0.017)
above	issue (0.013) below (0.015) level (0.015) point (0.016) during (0.017) end (0.017) day (0.017) price (0.017) final (0.017) over (0.017)
analyst	analysts (0.019) investor (0.020) an (0.020) at (0.020) firm (0.020) it (0.020) a (0.020) another (0.020) smith (0.020) investment (0.020)
base	on (0.012) near (0.012) in (0.012) of (0.012) which (0.012) through (0.012) the (0.012) line (0.012) also (0.012) huge (0.012)
close	down (0.014) friday (0.015) up (0.016) monday (0.016) heavy (0.016) tuesday (0.016) at (0.016) yesterday (0.017) stock (0.017) continued (0.017)
concern	group (0.010) based (0.011) maker (0.012) acquisition (0.012) including (0.013) mass (0.013) its (0.014) management (0.014) financial (0.014) held (0.014)
deal	without (0.010) that (0.010) way (0.010) run (0.010) but (0.011) saying (0.011) out (0.011) is (0.011) considered (0.011) having (0.011)
despite	in (0.007) while (0.008) on (0.008) to (0.009) after (0.009) fall (0.009) performance (0.009) of (0.009) the (0.009) industry (0.009)
even	so (0.006) now (0.006) still (0.007) but (0.007) just (0.007) once (0.008) that (0.008) yet (0.008) not (0.008) having (0.008)
expects	plans (0.022) boost (0.023) earnings (0.023) its (0.023) related (0.024) continuing (0.024) agreed (0.024) sell (0.024) declined (0.024) results (0.024)
family	public (0.012) called (0.012) great (0.012) office (0.012) made (0.012) with (0.012) through (0.012) into (0.012) out (0.012) of (0.012)
gained	dropped (0.024) fell (0.026) close (0.028) sell (0.029) agreed (0.029) declined (0.029) according (0.030) lost (0.030) plans (0.031) allow (0.031)
george	richard (0.011) michael (0.011) who (0.011) james (0.011) paul (0.011) robert (0.011) john (0.011) david (0.012) william (0.012) former (0.012)
germany	german (0.017) west (0.018) east (0.020) europe (0.021) japan (0.022) european (0.023) canada (0.023) world (0.023) growing (0.024) major (0.024)
general	finance (0.011) a (0.011) international (0.011) recently (0.011) and (0.012) financial (0.012) service (0.012) american (0.012) an (0.012) limited (0.012)
hard	find (0.008) get (0.008) too (0.009) them (0.009) how (0.009) say (0.009) go (0.009) having (0.010) make (0.010) come (0.010)
included	gain (0.013) nine (0.016) million (0.017) charge (0.017) results (0.018) continuing (0.018) reported (0.018) year (0.019) profit (0.019) period (0.019)
independent	outside (0.008) potential (0.009) control (0.010) also (0.010) is (0.010) the (0.010) public (0.011) that (0.011) through (0.011) under (0.011)
index	points (0.022) dow (0.026) stocks (0.026) volume (0.026) futures (0.026) monday (0.027) jones (0.027) traders (0.028) issues (0.028) traded (0.028)
it's	that's (0.008) there's (0.009) like (0.009) good (0.009) very (0.009) just (0.009) really (0.010) something (0.010) see (0.010) getting (0.010)
labor	washington (0.012) workers (0.013) by (0.013) department (0.013) and (0.013) in (0.014) major (0.014) the (0.014) industry (0.014) to (0.014)
making	with (0.006) now (0.007) and (0.007) to (0.007) another (0.007) that (0.007) is (0.007) taking (0.007) having (0.007) but (0.007)
men	women (0.011) children (0.015) these (0.015) groups (0.015) those (0.015) left (0.015) young (0.015) people (0.015) others (0.016) all (0.016)
night	then (0.011) when (0.011) around (0.012) turned (0.012) took (0.012) into (0.012) time (0.012) left (0.012) later (0.012) run (0.012)
nuclear	power (0.013) public (0.017) plant (0.017) itself (0.017) project (0.018) under (0.018) system (0.018) by (0.018) future (0.018) military (0.018)
old	former (0.013) who (0.013) richard (0.014) john (0.015) george (0.015) young (0.015) job (0.015) robert (0.016) paul (0.016) james (0.016)
operating	operations (0.015) division (0.019) services (0.020) post (0.020) financial (0.021) products (0.021) including (0.021) concern (0.021) continuing (0.021) results (0.021)

Table A.10 (contd.)

paid	for (0.011) about (0.011) full (0.012) which (0.012) of (0.012) a (0.012) in (0.012) also (0.012) the (0.012) an (0.013)
partners	management (0.011) firm (0.011) investment (0.011) limited (0.013) business (0.013) group (0.013) a (0.013) including (0.013) seeking (0.013) held (0.013)
percentage	point (0.022) five (0.023) rate (0.023) last (0.023) slightly (0.023) average (0.023) seven (0.023) change (0.023) above (0.023) jan (0.024)
political	leaders (0.009) democratic (0.010) country (0.010) left (0.010) press (0.011) policy (0.011) only (0.011) toward (0.011) best (0.011) important (0.011)
preferred	common (0.016) holders (0.022) shares (0.023) outstanding (0.024) purchase (0.024) offering (0.026) receive (0.027) cash (0.028) each (0.028) dividend (0.029)
product	line (0.008) with (0.010) for (0.010) also (0.010) making (0.010) in (0.010) to (0.011) the (0.011) is (0.011) potential (0.011)
products	businesses (0.011) services (0.012) business (0.012) food (0.012) lines (0.012) mass (0.012) equipment (0.013) marketing (0.013) international (0.013) resources (0.013)
same	during (0.012) final (0.013) process (0.013) government (0.013) level (0.013) economy (0.013) effect (0.014) nation's (0.014) act (0.014) thus (0.014)
should	must (0.007) could (0.008) might (0.010) can (0.011) may (0.011) not (0.011) won't (0.011) would (0.011) that (0.011) yet (0.011)
take	make (0.006) give (0.007) help (0.007) be (0.008) keep (0.009) come (0.009) go (0.009) turn (0.009) start (0.009) move (0.009)
they're	don't (0.011) aren't (0.011) see (0.011) that's (0.012) people (0.012) it's (0.012) doing (0.012) like (0.012) getting (0.012) they (0.012)
times	on (0.011) in (0.011) of (0.011) almost (0.011) hit (0.011) around (0.011) over (0.011) time (0.012) at (0.012) two (0.012)
transaction	company (0.012) purchase (0.014) proposed (0.014) company's (0.015) merger (0.015) agreement (0.015) sale (0.015) planned (0.015) acquisition (0.016) completed (0.016)
transportation	off (0.015) up (0.020) industrial (0.022) financial (0.022) service (0.022) general (0.022) international (0.022) and (0.022) washington (0.022) markets (0.022)
use	used (0.009) help (0.010) using (0.010) lead (0.010) work (0.010) support (0.011) take (0.011) all (0.011) potential (0.011) be (0.011)
using	making (0.008) run (0.008) without (0.008) into (0.008) to (0.008) that (0.008) free (0.008) is (0.008) but (0.008) outside (0.009)
wall	journal (0.025) street (0.026) offers (0.027) staff (0.028) publishing (0.029) acquisitions (0.030) mergers (0.031) news (0.031) europ (0.032) stk (0.032)
wednesday	tuesday (0.010) monday (0.012) friday (0.015) late (0.016) down (0.017) slightly (0.018) heavy (0.019) up (0.019) close (0.020) average (0.020)
week	month (0.013) last (0.015) after (0.015) fall (0.015) in (0.016) on (0.016) to (0.016) meanwhile (0.016) while (0.016) at (0.016)
weren't	terms (0.015) held (0.015) its (0.015) it (0.016) also (0.016) said (0.016) disclosed (0.016) an (0.016) planned (0.016) business (0.016)
will	may (0.010) be (0.011) future (0.011) also (0.011) to (0.011) won't (0.011) remain (0.011) used (0.012) through (0.012) today (0.012)
william	robert (0.005) richard (0.005) john (0.006) james (0.006) david (0.007) paul (0.008) michael (0.009) jr (0.011) director (0.012) george (0.012)
workers	employees (0.011) jobs (0.011) labor (0.013) work (0.013) local (0.013) other (0.013) two (0.014) several (0.014) all (0.014) are (0.014)

Figure A.10 below shows the dendrograms containing the 1000 target words considered in analysis 10 of Chapter 4.

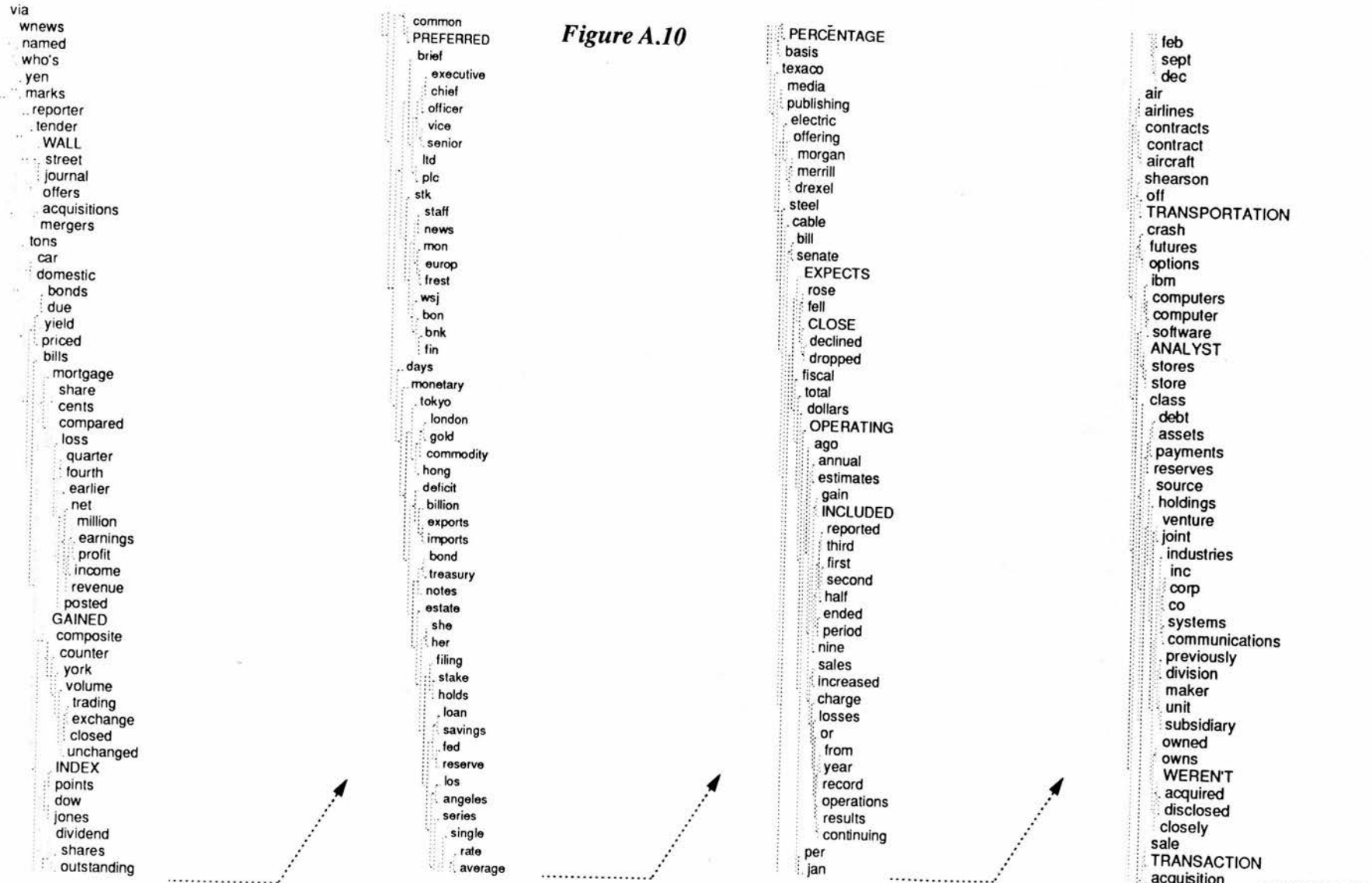

via
wnews
named
who's
yen
marks
reporter
tender
WALL
street
journal
offers
acquisitions
mergers
tons
car
domestic
bonds
due
yield
priced
bills
mortgage
share
cents
compared
loss
quarter
fourth
earlier
net
million
earnings
profit
income
revenue
posted
GAINED
composite
counter
york
volume
trading
exchange
closed
unchanged
INDEX
points
dow
jones
dividend
shares
outstanding
common
PREFERRED
brief
executive
chief
officer
vice
senior
ltd
plc
stk
staff
news
mon
europ
frest
wsj
bon
bnk
fin
days
monetary
tokyo
london
gold
commodity
hong
deficit
billion
exports
imports
bond
treasury
notes
estate
she
her
filing
stake
holds
loan
savings
fed
reserve
los
angeles
series
single
rate
average
PERCENTAGE
basis
texaco
media
publishing
electric
offering
morgan
merrill
drexel
steel
cable
bill
senate
EXPECTS
rose
fell
CLOSE
declined
dropped
fiscal
total
dollars
OPERATING
ago
annual
estimates
gain
INCLUDED
reported
third
first
second
half
ended
period
nine
sales
increased
charge
losses
or
from
year
record
operations
results
continuing
per
jan
feb
sept
dec
air
airlines
contracts
contract
aircraft
shearson
off
TRANSPORTATION
crash
futures
options
ibm
computers
computer
software
ANALYST
stores
store
class
debt
assets
payments
reserves
source
holdings
venture
joint
industries
inc
corp
co
systems
communications
previously
division
maker
unit
subsidiary
owned
owns
WEREN'T
acquired
disclosed
closely
sale
TRANSACTION
acquisition

Figure A.10

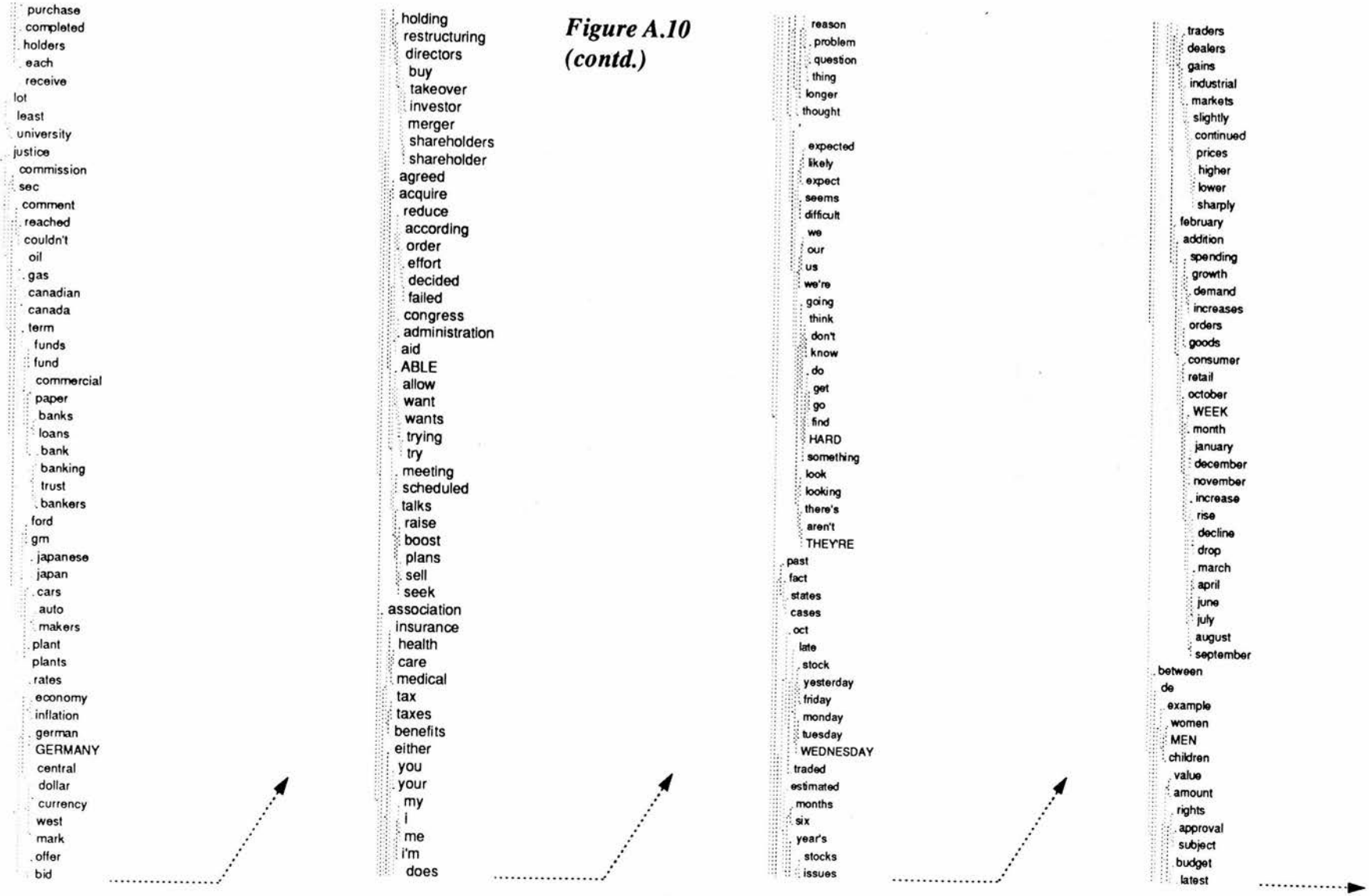


Figure A.10 (contd.)

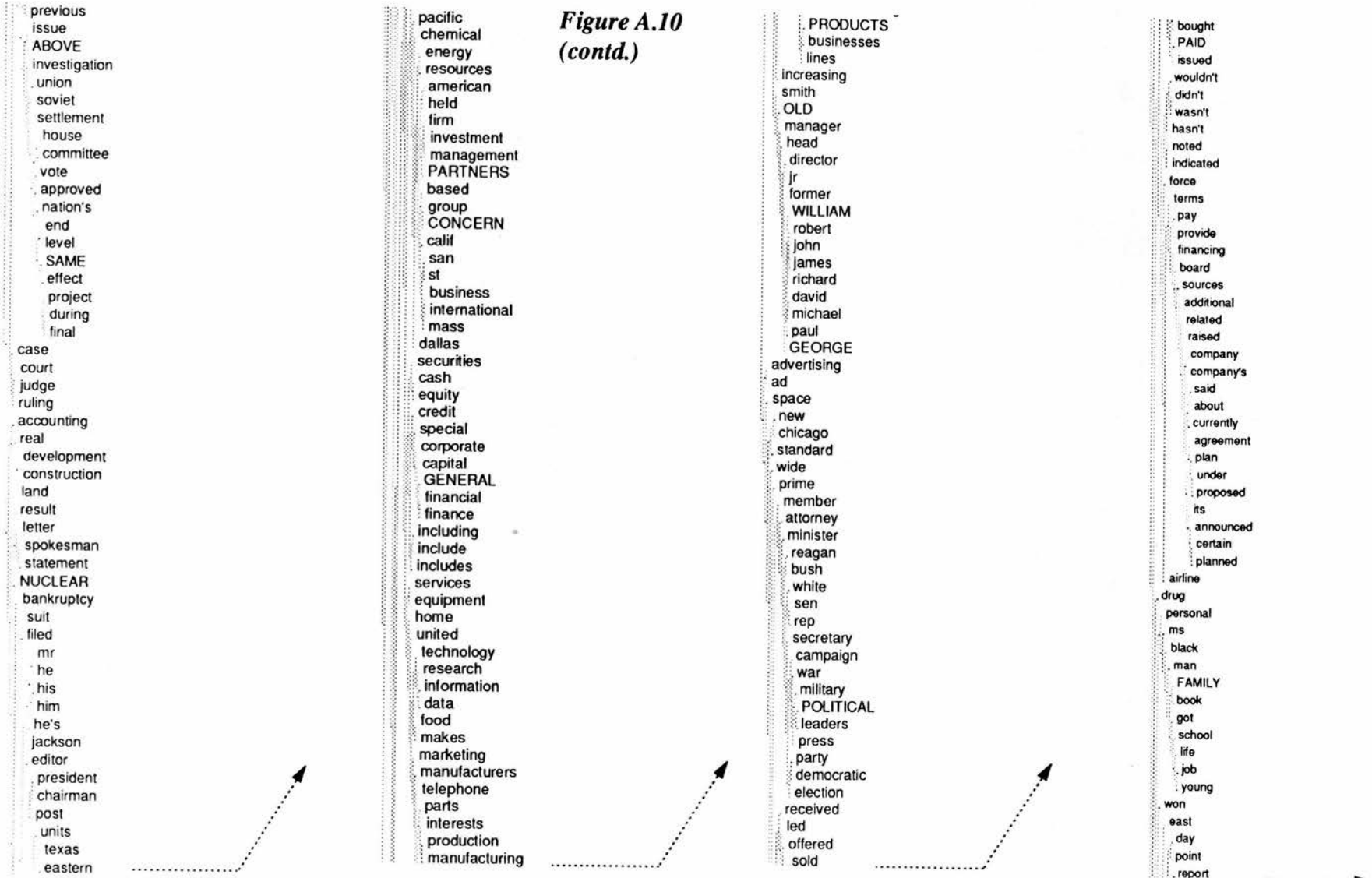


Figure A.10 (contd.)

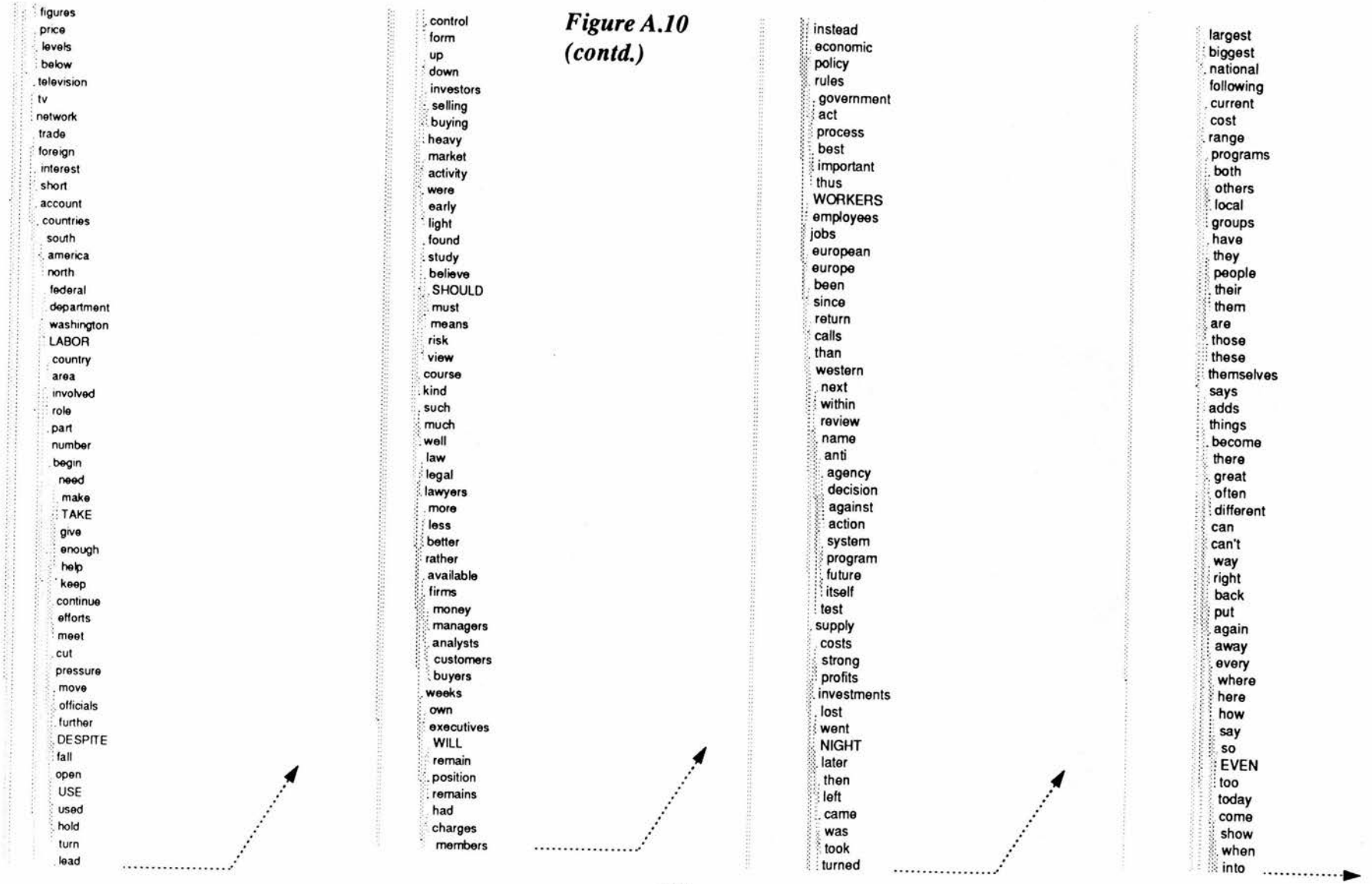


Figure A.10 (contd.)

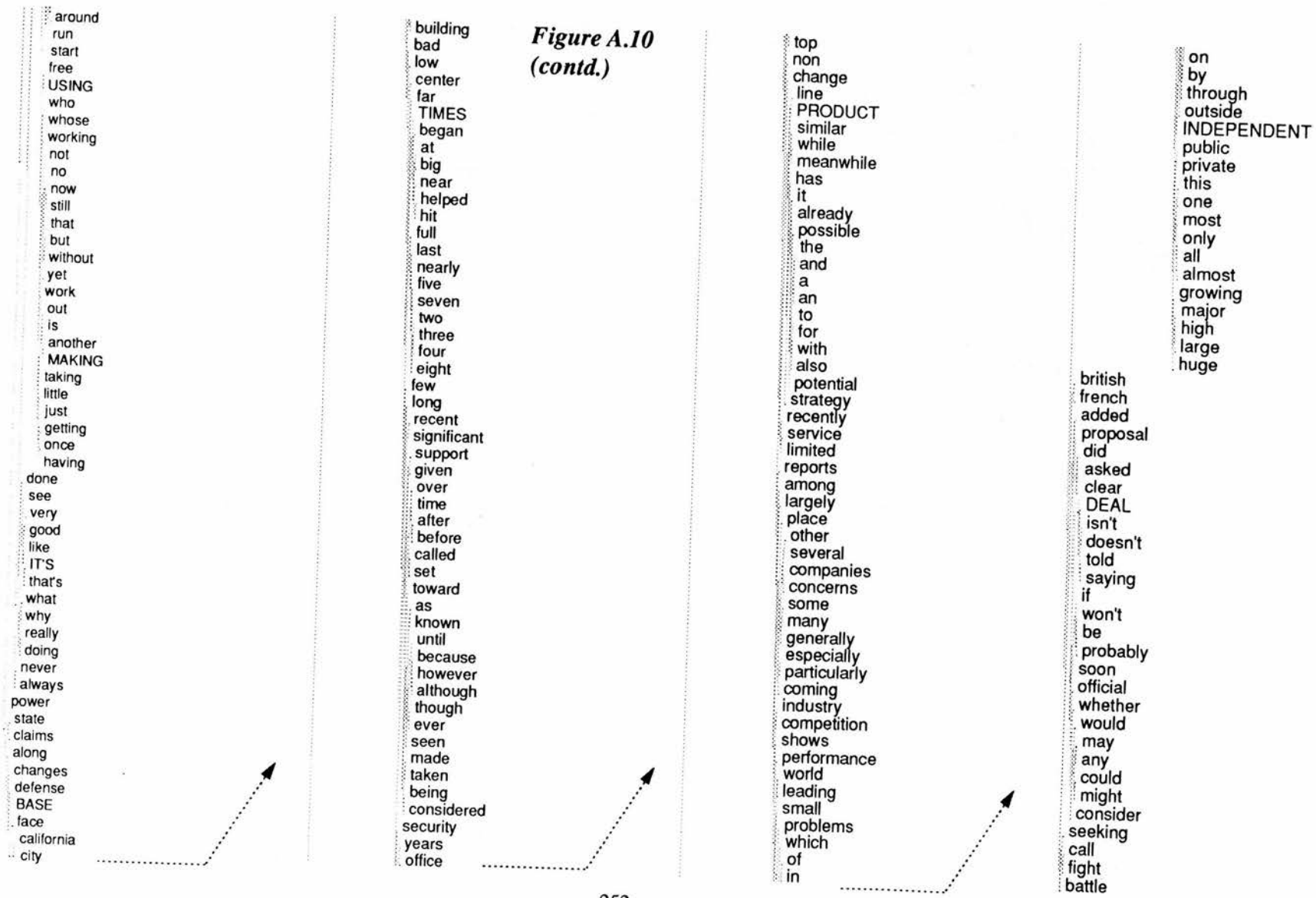


Figure A.10 (contd.)

Table A.11
(Spearman Distance Metric, Window Length=100)

The table below contains the 50 target words and 10 nearest neighbours for analysis 11 in Chapter 4.

Target Word	10 Nearest Neighbours (Spearman Correlation Coefficient)
able	have (0.943) make (0.941) such (0.939) if (0.935) could (0.934) be (0.933) take (0.931) that (0.930) they (0.929) but (0.929)
above	higher (0.912) below (0.910) slightly (0.902) points (0.902) at (0.900) lower (0.899) down (0.899) average (0.899) point (0.898) markets (0.898)
analyst	analysts (0.950) selling (0.895) buying (0.886) market (0.885) profits (0.882) big (0.882) down (0.880) up (0.879) heavy (0.879) expect (0.875)
base	near (0.867) large (0.863) low (0.862) high (0.862) next (0.858) than (0.856) more (0.855) in (0.853) end (0.853) past (0.852)
close	friday (0.929) yesterday (0.926) closed (0.923) at (0.917) down (0.917) added (0.916) trading (0.914) up (0.912) stock (0.912) end (0.908)
concern	based (0.959) previously (0.949) company (0.948) said (0.948) inc (0.945) holding (0.945) unit (0.942) acquisition (0.939) group (0.936) its (0.936)
deal	out (0.914) any (0.912) be (0.907) isn't (0.906) if (0.905) own (0.904) take (0.903) has (0.902) doesn't (0.899) might (0.898)
despite	while (0.953) recent (0.943) up (0.942) than (0.939) strong (0.937) since (0.936) fall (0.936) down (0.936) in (0.935) helped (0.935)
even	so (0.983) now (0.974) only (0.974) too (0.973) though (0.971) they (0.971) there (0.969) come (0.967) just (0.966) but (0.964)
expects	net (0.944) cents (0.940) quarter (0.935) said (0.931) loss (0.930) million (0.929) earnings (0.928) fourth (0.928) reported (0.924) profit (0.924)
family	own (0.875) life (0.874) city (0.871) whose (0.869) all (0.865) it (0.864) name (0.862) with (0.862) has (0.862) who (0.861)
gained	volume (0.914) fell (0.910) rose (0.899) dropped (0.896) continued (0.881) sharply (0.881) trading (0.880) heavy (0.878) closed (0.877) dow (0.877)
george	who (0.924) he (0.922) mr (0.908) his (0.908) richard (0.898) left (0.897) this (0.891) no (0.890) whose (0.890) old (0.889)
germany	west (0.898) german (0.894) europe (0.876) japan (0.876) european (0.871) foreign (0.861) world (0.846) central (0.842) countries (0.829) frest (0.825)
general	of (0.906) by (0.903) co (0.902) for (0.902) a (0.901) two (0.901) including (0.899) new (0.899) which (0.898) three (0.898)
hard	so (0.957) too (0.953) like (0.953) even (0.950) now (0.949) way (0.948) just (0.948) what (0.948) it's (0.947) do (0.947)
included	related (0.914) continuing (0.914) from (0.912) million (0.911) gain (0.910) reported (0.909) loss (0.907) latest (0.905) third (0.904) net (0.902)
independent	control (0.894) has (0.891) seek (0.889) seeking (0.889) under (0.888) it (0.887) wouldn't (0.886) any (0.885) decision (0.885) hasn't (0.880)
index	volume (0.924) stocks (0.917) prices (0.912) traders (0.908) market (0.905) points (0.901) lower (0.900) higher (0.898) fell (0.898) sharply (0.898)
it's	just (0.975) so (0.974) too (0.971) like (0.970) think (0.969) don't (0.968) do (0.968) what (0.967) get (0.966) that's (0.966)
labor	workers (0.891) the (0.863) number (0.862) for (0.860) in (0.856) changes (0.856) benefits (0.856) to (0.854) department (0.854) nation's (0.853)
making	make (0.958) has (0.953) that (0.953) have (0.953) other (0.953) is (0.950) but (0.950) more (0.948) into (0.947) with (0.947)
men	young (0.906) people (0.904) who (0.903) man (0.902) women (0.901) never (0.899) his (0.898) him (0.897) her (0.893) know (0.892)
night	his (0.889) place (0.881) him (0.874) he (0.872) then (0.871) who (0.871) never (0.870) out (0.869) i (0.867) left (0.865)
nuclear	power (0.856) defense (0.838) military (0.829) state (0.804) process (0.803) force (0.801) official (0.800) project (0.797) would (0.797) anti (0.794)
old	who (0.944) he (0.915) left (0.914) mr (0.911) years (0.907) his (0.906) become (0.901) whose (0.898) name (0.893) i (0.893)
operating	operations (0.944) net (0.924) expects (0.920) officer (0.915) company's (0.915) company (0.908) loss (0.908) revenue (0.905) profit (0.903) business (0.901)

Table A.11 (contd.)

paid	pay (0.936) or (0.895) for (0.894) certain (0.890) a (0.888) by (0.888) of (0.887) it (0.884) payments (0.884) under (0.883)
partners	owns (0.884) agreed (0.883) comment (0.882) held (0.882) group (0.878) stake (0.877) firm (0.877) company (0.876) owned (0.872) reached (0.870)
percentage	average (0.908) rate (0.907) higher (0.907) lower (0.901) year (0.895) slightly (0.893) month (0.883) below (0.883) above (0.880) compared (0.878)
political	democratic (0.948) party (0.935) leaders (0.926) reagan (0.918) bush (0.906) white (0.902) war (0.899) election (0.892) campaign (0.887) sen (0.886)
preferred	common (0.943) holders (0.933) outstanding (0.930) purchase (0.910) tender (0.902) transaction (0.899) acquisition (0.897) sale (0.896) shares (0.893) acquire (0.893)
product	line (0.894) research (0.891) industry (0.891) computers (0.874) makers (0.874) products (0.873) marketing (0.872) competition (0.871) technology (0.869) computer (0.867)
products	business (0.920) maker (0.914) equipment (0.910) its (0.907) based (0.897) said (0.896) company (0.896) industries (0.896) division (0.892) operating (0.892)
same	only (0.961) more (0.960) have (0.957) but (0.956) all (0.956) those (0.952) are (0.948) still (0.947) this (0.947) now (0.947)
should	have (0.960) not (0.958) if (0.957) such (0.956) that (0.953) those (0.951) without (0.949) can (0.948) only (0.947) they (0.947)
take	that (0.966) have (0.964) is (0.960) but (0.959) be (0.958) out (0.958) if (0.957) could (0.955) has (0.954) make (0.954)
they're	it's (0.955) lot (0.951) don't (0.949) that's (0.947) just (0.946) says (0.944) there's (0.943) too (0.941) think (0.941) going (0.939)
times	time (0.910) most (0.905) when (0.898) around (0.898) but (0.896) one (0.895) some (0.894) another (0.894) well (0.893) good (0.891)
transaction	acquire (0.952) acquisition (0.951) tender (0.941) purchase (0.936) stake (0.933) acquired (0.931) disclosed (0.924) outstanding (0.923) sale (0.921) completed (0.920)
transportation	said (0.857) which (0.856) including (0.854) plans (0.850) staff (0.846) its (0.846) announced (0.844) also (0.841) additional (0.840) by (0.839)
use	used (0.934) using (0.930) such (0.924) help (0.912) can (0.912) is (0.910) called (0.910) have (0.908) be (0.905) make (0.902)
using	use (0.930) such (0.920) can (0.916) have (0.912) example (0.909) them (0.907) their (0.907) they (0.906) only (0.905) get (0.904)
wall	street (0.998) journal (0.985) reporter (0.931) news (0.922) staff (0.921) new (0.919) financial (0.908) york (0.904) said (0.901) international (0.899)
wednesday	monday (0.928) friday (0.928) tuesday (0.918) late (0.896) trading (0.896) markets (0.891) yesterday (0.887) unchanged (0.881) exchange (0.879) traded (0.876)
week	month (0.919) previous (0.916) last (0.904) at (0.897) down (0.894) day (0.893) from (0.892) april (0.892) three (0.891) march (0.889)
weren't	business (0.926) based (0.924) its (0.924) said (0.922) which (0.915) disclosed (0.913) closely (0.913) co (0.913) held (0.913) unit (0.912)
will	of (0.953) also (0.950) to (0.950) for (0.949) and (0.948) be (0.946) which (0.946) a (0.944) about (0.944) new (0.943)
william	john (0.941) robert (0.935) david (0.927) richard (0.924) james (0.922) president (0.920) director (0.913) and (0.906) an (0.905) paul (0.905)
workers	jobs (0.893) labor (0.891) work (0.883) force (0.856) benefits (0.856) employees (0.854) cost (0.851) working (0.846) same (0.843) only (0.841)

Figure A.11 below shows the dendrogram containing the 1000 target words considered in analysis 11 of Chapter 4.

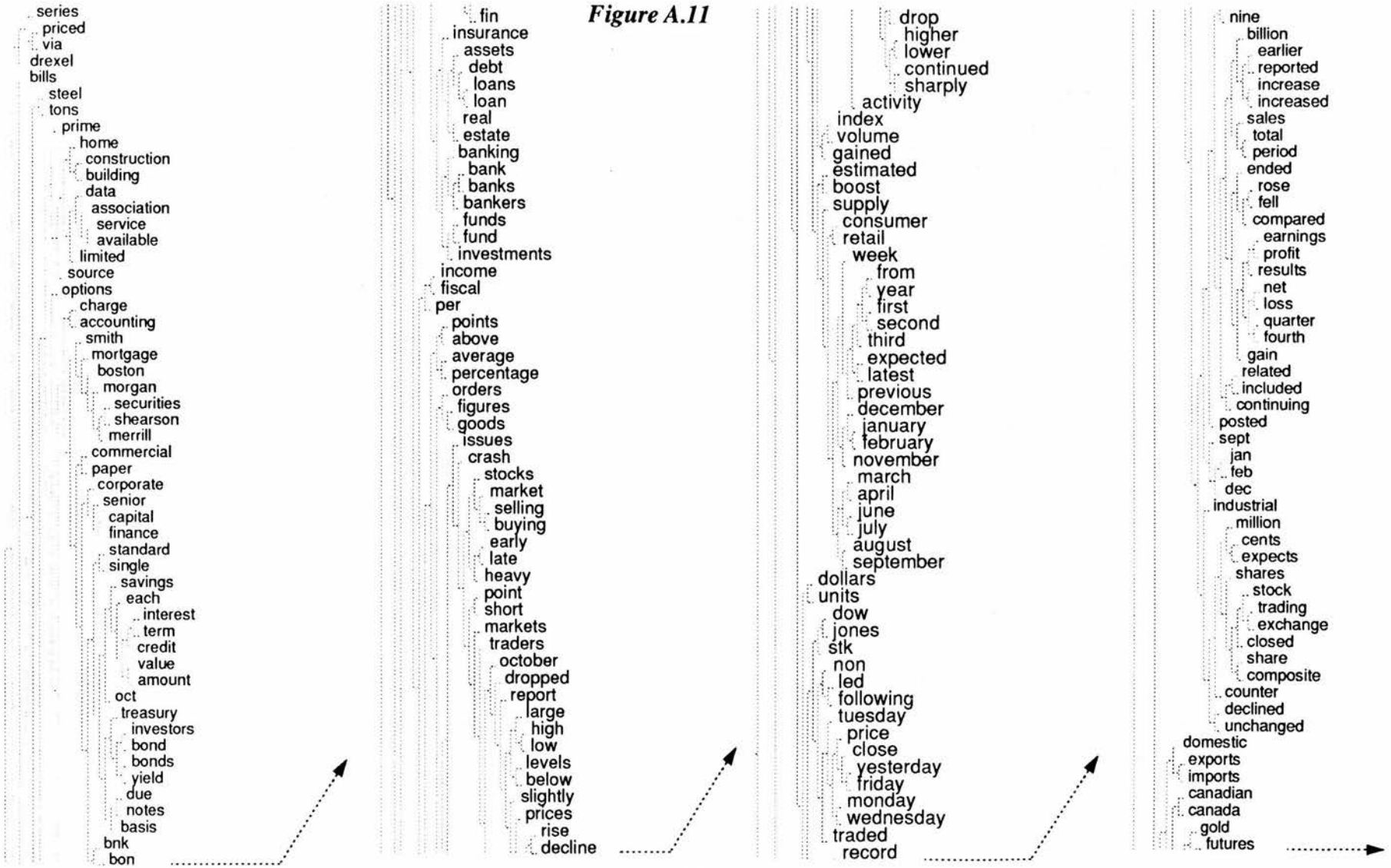

series
priced
via
drexel
bills
steel
tons
prime
home
construction
building
data
association
service
available
limited
source
options
charge
accounting
smith
mortgage
boston
morgan
securities
shearson
merrill
commercial
paper
corporate
senior
capital
finance
standard
single
savings
each
interest
term
credit
value
amount
oct
treasury
investors
bond
bonds
yield
due
notes
basis
bnk
bon
fin
insurance
assets
debt
loans
loan
real
estate
banking
bank
banks
bankers
funds
fund
investments
income
fiscal
per
points
above
average
percentage
orders
figures
goods
issues
crash
stocks
market
selling
buying
early
late
heavy
point
short
markets
traders
october
dropped
report
large
high
low
levels
below
slightly
prices
rise
decline
drop
higher
lower
continued
sharply
activity
index
volume
gained
estimated
boost
supply
consumer
retail
week
from
year
first
second
third
expected
latest
previous
december
january
february
november
march
april
june
july
august
september
dollars
units
dow
jones
stk
non
led
following
tuesday
price
close
yesterday
friday
monday
wednesday
traded
record
nine
billion
earlier
reported
increase
increased
sales
total
period
ended
rose
fell
compared
earnings
profit
results
net
loss
quarter
fourth
gain
related
included
continuing
posted
sept
jan
feb
dec
industrial
million
cents
expects
shares
stock
trading
exchange
closed
share
composite
counter
declined
unchanged
domestic
exports
imports
canadian
canada
gold
futures

Figure A.11

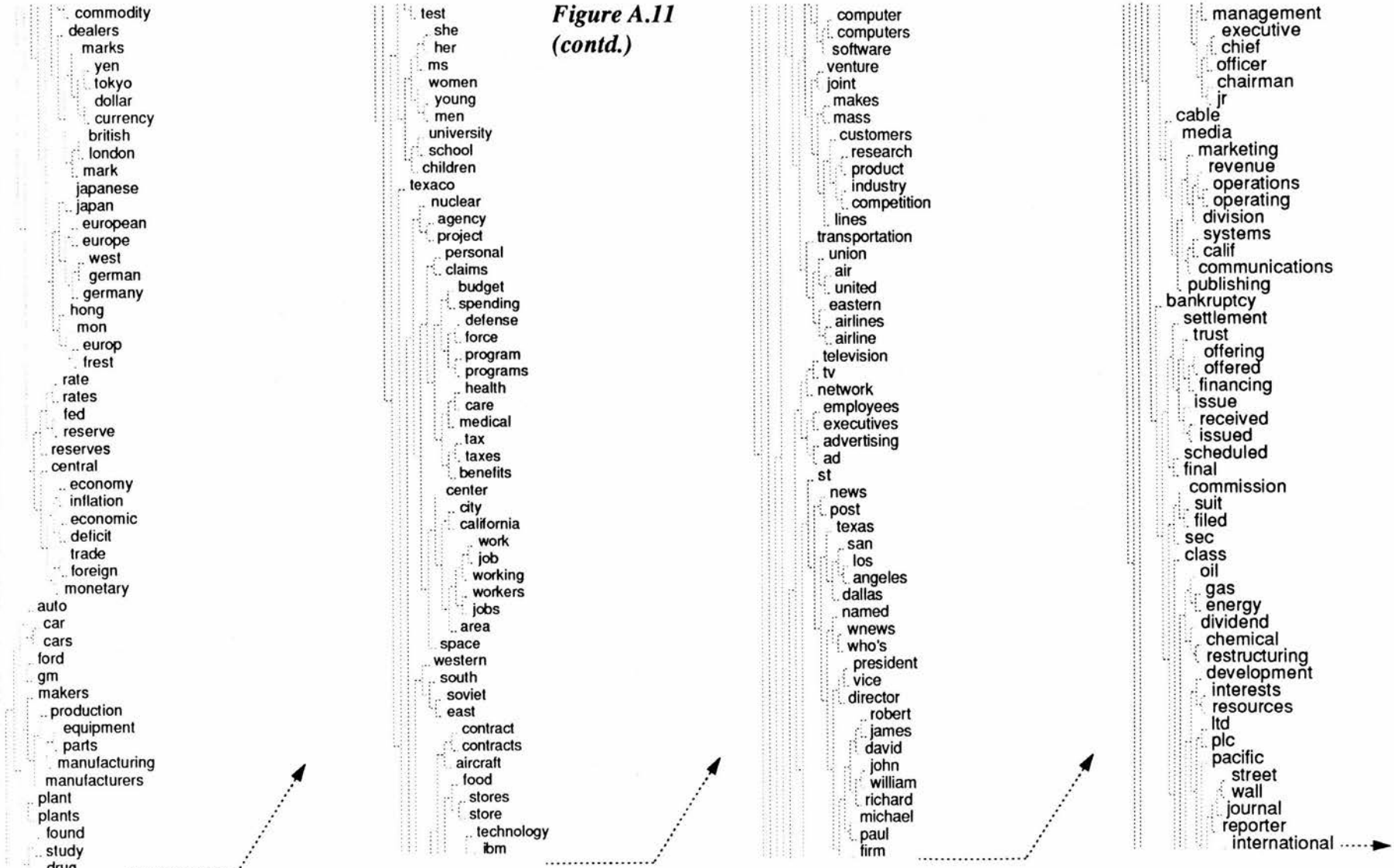


Figure A.11 (contd.)

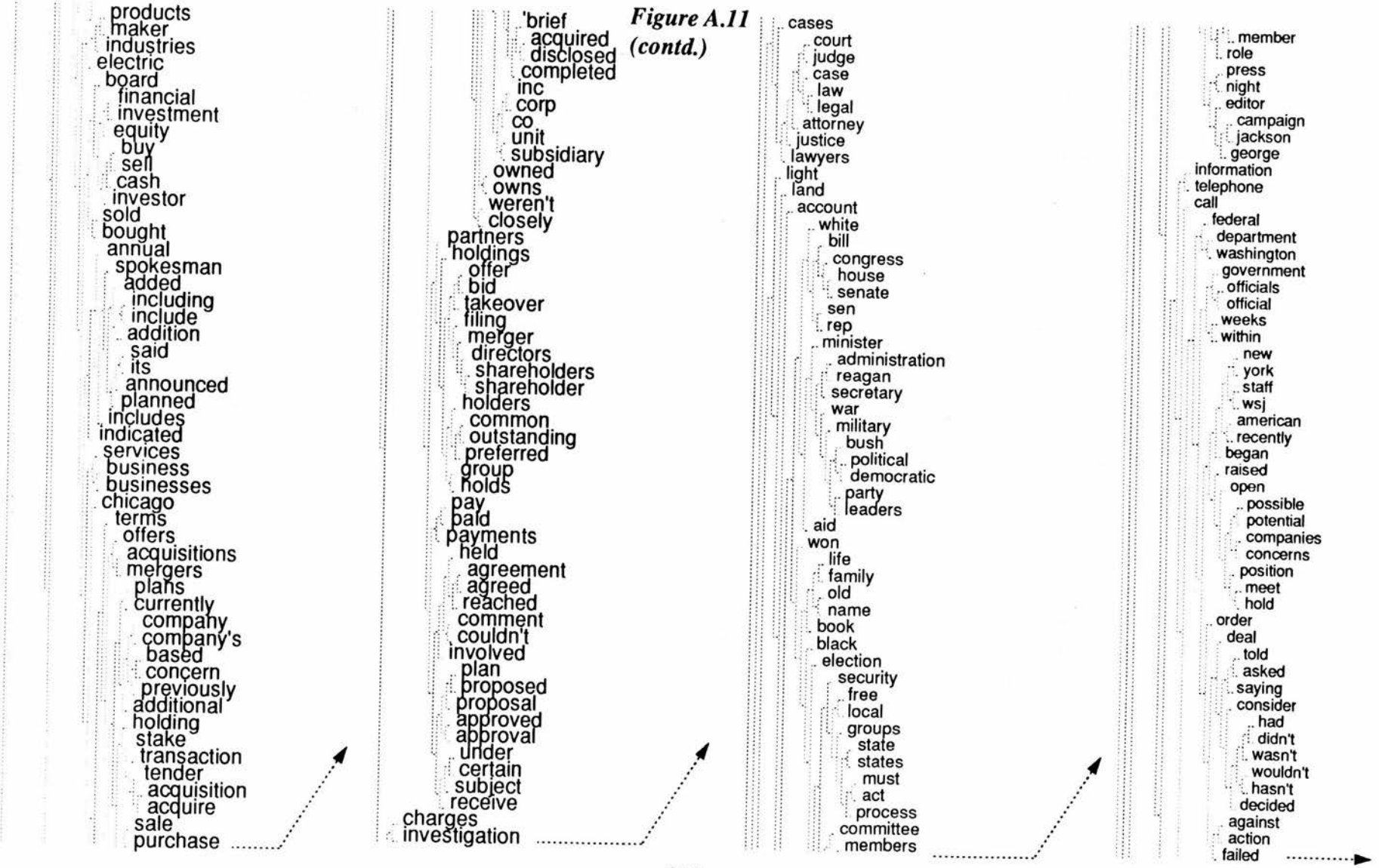


Figure A.11 (contd.)

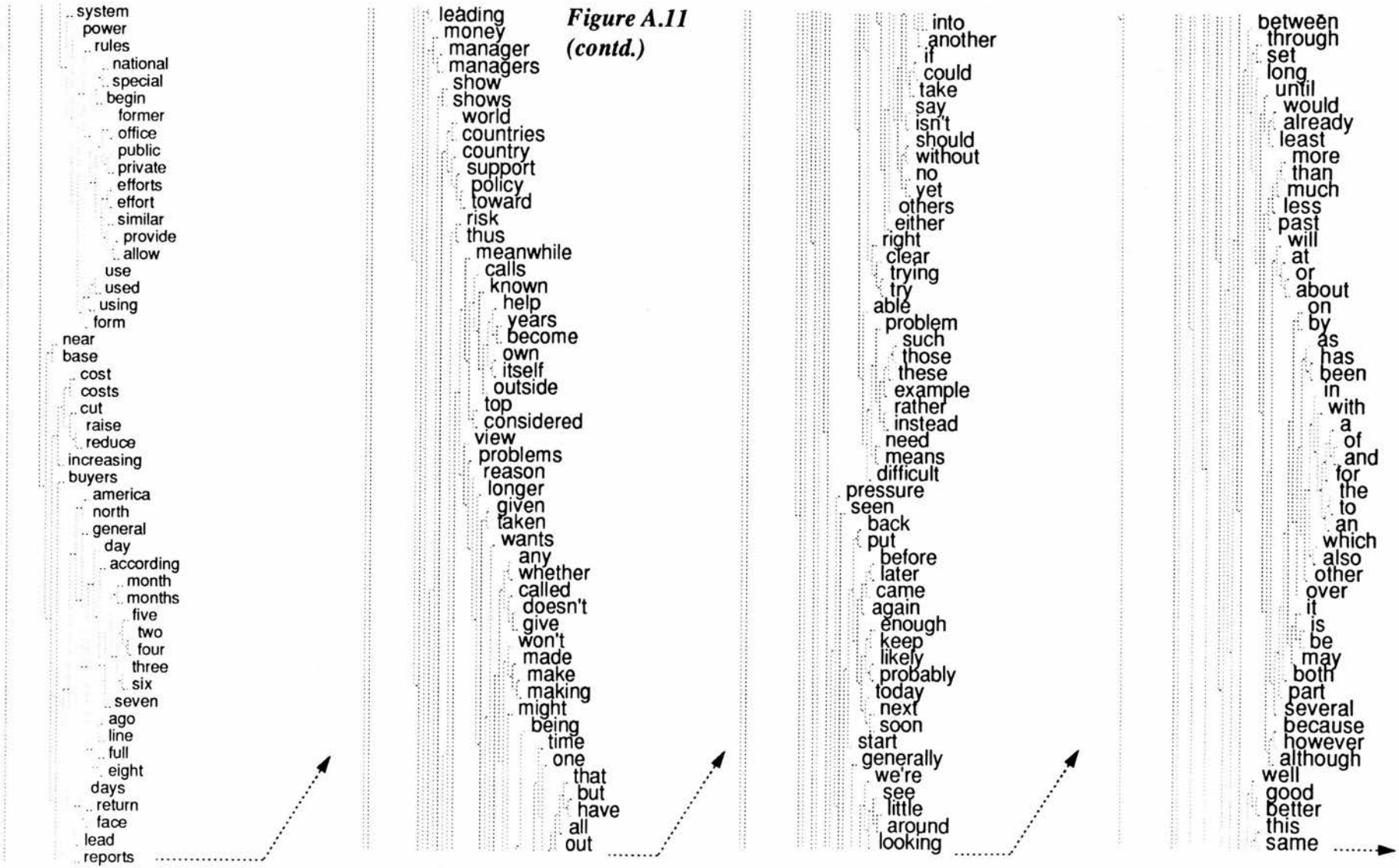


Figure A.11 (contd.)

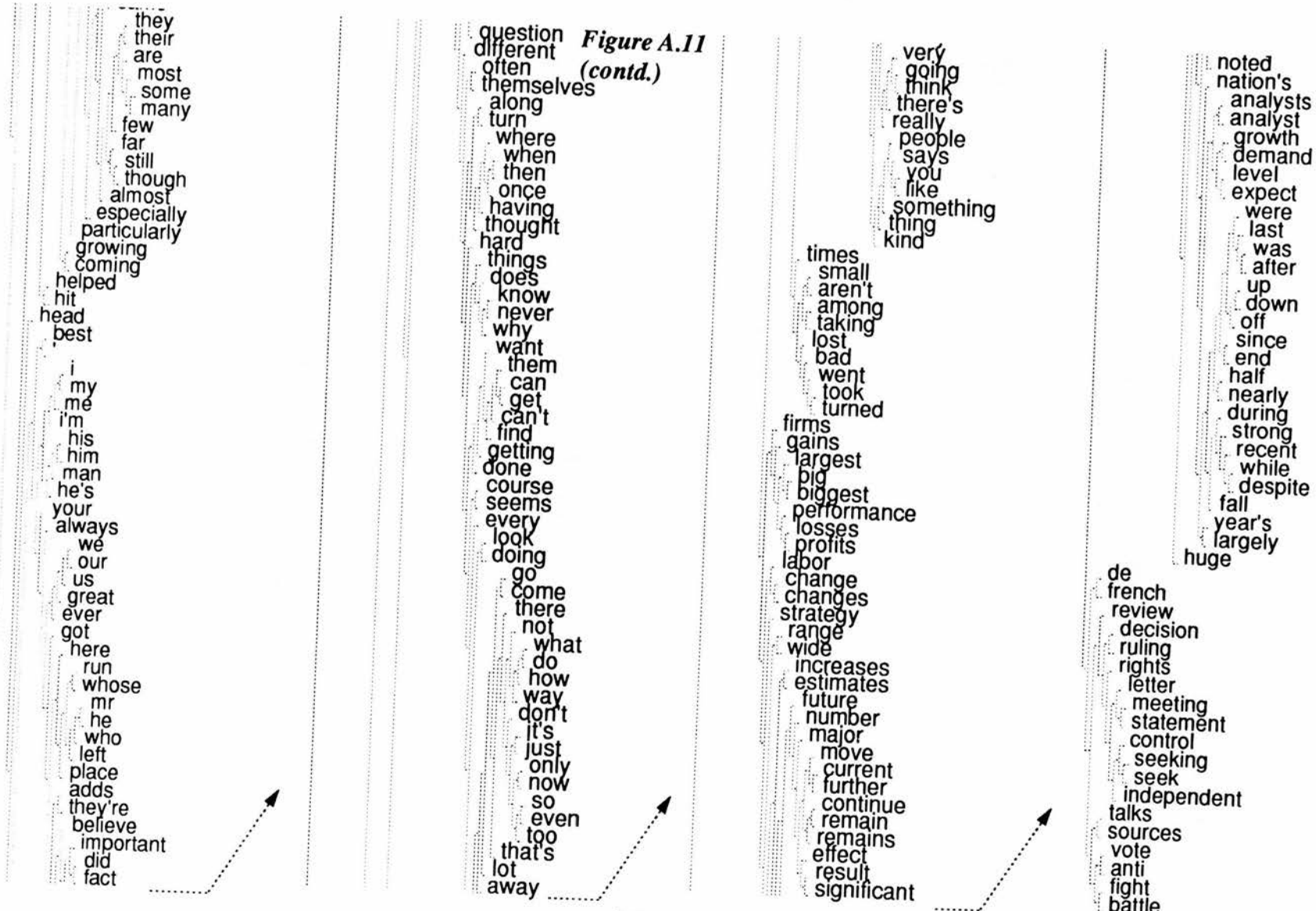


Figure A.11 (contd.)

Table A.12
(Euclidean Distance Metric, Window Length=100)

The table below contains the 50 target words and 10 nearest neighbours for analysis 12 in Chapter 4.

Target Word	10 Nearest Neighbours (Euclidean Distance)
able	likely (0.006) take (0.006) keep (0.006) be (0.006) difficult (0.006) trying (0.006) enough (0.007) give (0.007) try (0.007) make (0.007)
above	issue (0.008) point (0.009) led (0.009) basis (0.009) end (0.009) below (0.009) were (0.010) by (0.010) following (0.010) seven (0.010)
analyst	analysts (0.006) boost (0.010) at (0.010) indicated (0.010) selling (0.010) added (0.010) said (0.010) up (0.010) yesterday (0.010) buy (0.010)
base	on (0.006) near (0.006) the (0.006) while (0.007) to (0.007) in (0.007) by (0.007) of (0.007) through (0.007) which (0.007)
close	heavy (0.005) down (0.006) friday (0.006) yesterday (0.007) at (0.007) up (0.007) price (0.008) which (0.008) selling (0.008) after (0.008)
concern	based (0.005) calif (0.005) previously (0.005) maker (0.005) company (0.006) disclosed (0.006) acquired (0.006) industries (0.006) unit (0.006) closely (0.007)
deal	put (0.005) out (0.005) take (0.005) saying (0.005) that (0.005) might (0.005) isn't (0.005) could (0.005) any (0.005) but (0.005)
despite	while (0.003) in (0.004) nearly (0.004) largely (0.004) recent (0.004) last (0.004) after (0.005) fall (0.005) on (0.005) up (0.005)
even	so (0.002) now (0.003) though (0.003) too (0.003) way (0.003) only (0.004) come (0.004) just (0.004) how (0.004) no (0.004)
expects	million (0.010) operations (0.010) results (0.010) restructuring (0.011) continuing (0.011) reported (0.011) earnings (0.011) company (0.011) company's (0.012) sale (0.012)
family	whose (0.007) with (0.007) outside (0.007) out (0.008) known (0.008) and (0.008) into (0.008) made (0.008) all (0.008) own (0.008)
gained	dropped (0.015) fell (0.017) close (0.019) unchanged (0.019) volume (0.019) heavy (0.019) continued (0.019) declined (0.020) lost (0.020) issues (0.020)
george	who (0.005) michael (0.006) richard (0.006) paul (0.006) head (0.006) become (0.006) whose (0.006) david (0.006) top (0.007) best (0.007)
germany	west (0.008) german (0.010) europe (0.011) european (0.011) east (0.012) major (0.012) world (0.012) growing (0.012) japan (0.013) change (0.013)
general	including (0.006) mass (0.007) recently (0.007) a (0.007) for (0.007) development (0.007) and (0.007) include (0.007) held (0.007) also (0.007)
hard	too (0.004) how (0.004) get (0.004) find (0.004) now (0.004) so (0.004) just (0.004) getting (0.004) away (0.004) them (0.004)
included	gain (0.007) continuing (0.007) results (0.007) reported (0.008) nine (0.009) period (0.009) latest (0.009) earlier (0.009) charge (0.010) third (0.010)
independent	outside (0.005) control (0.005) potential (0.005) made (0.005) also (0.006) an (0.006) with (0.006) has (0.006) possible (0.006) already (0.006)
index	points (0.011) stocks (0.012) volume (0.012) traders (0.013) activity (0.014) monday (0.014) traded (0.014) issues (0.015) dow (0.015) prices (0.015)
it's	that's (0.003) there's (0.003) just (0.004) getting (0.004) really (0.004) can't (0.004) look (0.004) like (0.004) get (0.004) too (0.004)
labor	workers (0.007) washington (0.008) the (0.008) department (0.008) by (0.008) set (0.008) in (0.008) nation's (0.008) changes (0.008) of (0.008)
making	with (0.003) another (0.003) but (0.003) has (0.003) other (0.003) out (0.004) is (0.004) and (0.004) outside (0.004) as (0.004)
men	young (0.008) every (0.009) people (0.009) women (0.009) who (0.009) left (0.009) great (0.009) where (0.009) here (0.009) place (0.009)
night	then (0.006) when (0.006) turned (0.006) later (0.006) around (0.007) once (0.007) along (0.007) left (0.007) out (0.007) run (0.007)
nuclear	power (0.011) final (0.012) effect (0.012) project (0.012) anti (0.012) action (0.012) process (0.012) review (0.013) itself (0.013) defense (0.013)
old	john (0.007) paul (0.007) who (0.007) richard (0.007) george (0.007) head (0.008) former (0.008) job (0.008) michael (0.008) robert (0.008)
operating	operations (0.007) division (0.011) results (0.011) maker (0.011) continuing (0.011) nine (0.011) dec (0.011) reported (0.011) products (0.011) company's (0.011)

Table A.12 (contd.)

paid	for (0.006) pay (0.006) to (0.006) a (0.006) about (0.006) an (0.006) also (0.007) special (0.007) return (0.007) which (0.007)
partners	firm (0.008) held (0.008) management (0.008) business (0.008) including (0.008) owned (0.008) seeking (0.008) interests (0.008) group (0.008) owns (0.008)
percentage	higher (0.011) month (0.011) lower (0.011) slightly (0.011) average (0.012) below (0.012) light (0.012) five (0.012) term (0.012) low (0.012)
political	leaders (0.006) democratic (0.006) party (0.007) election (0.007) course (0.008) country (0.008) important (0.008) war (0.008) role (0.008) question (0.008)
preferred	common (0.009) holders (0.011) outstanding (0.011) shares (0.012) purchase (0.013) transaction (0.014) acquisition (0.015) shareholders (0.015) brief (0.015) completed (0.015)
product	line (0.005) industry (0.007) for (0.007) a (0.007) makes (0.007) also (0.007) an (0.007) it (0.007) and (0.007) about (0.007)
products	business (0.006) calif (0.007) industries (0.007) services (0.007) based (0.007) businesses (0.007) equipment (0.007) international (0.007) co (0.007) concern (0.008)
same	only (0.004) toward (0.005) taken (0.005) however (0.005) most (0.005) particularly (0.005) that (0.005) far (0.005) before (0.005) almost (0.005)
should	must (0.004) not (0.005) believe (0.005) could (0.005) means (0.005) clear (0.005) that (0.005) problem (0.005) only (0.005) without (0.005)
take	give (0.003) make (0.003) help (0.004) be (0.004) put (0.004) but (0.004) soon (0.004) another (0.004) keep (0.004) any (0.004)
they're	don't (0.004) that's (0.005) getting (0.005) there's (0.005) it's (0.005) say (0.005) look (0.005) get (0.006) doing (0.006) see (0.006)
times	time (0.006) around (0.006) almost (0.006) still (0.006) little (0.006) as (0.006) most (0.006) seen (0.006) into (0.006) hit (0.006)
transaction	purchase (0.006) completed (0.007) acquisition (0.007) disclosed (0.008) company (0.008) sale (0.009) merger (0.009) acquired (0.009) shareholders (0.009) terms (0.009)
transportation	staff (0.009) off (0.009) general (0.009) including (0.009) lines (0.009) international (0.009) development (0.009) service (0.009) include (0.009) special (0.009)
use	used (0.005) using (0.005) be (0.005) instead (0.006) help (0.006) itself (0.006) have (0.006) such (0.006) is (0.006) similar (0.006)
using	without (0.004) but (0.004) into (0.004) have (0.004) now (0.005) being (0.005) such (0.005) out (0.005) run (0.005) is (0.005)
wall	journal (0.006) street (0.006) reporter (0.008) staff (0.010) offers (0.010) news (0.010) mergers (0.011) publishing (0.011) international (0.011) financial (0.011)
wednesday	tuesday (0.006) monday (0.007) friday (0.009) late (0.010) slightly (0.010) dealers (0.011) traded (0.011) volume (0.011) sharply (0.011) continued (0.011)
week	last (0.008) month (0.008) meanwhile (0.008) weeks (0.009) at (0.009) day (0.009) up (0.009) on (0.009) while (0.009) to (0.009)
weren't	held (0.006) closely (0.006) terms (0.006) announced (0.006) including (0.007) said (0.007) its (0.007) business (0.007) san (0.007) mass (0.007)
will	remain (0.005) currently (0.005) be (0.006) begin (0.006) also (0.006) next (0.006) available (0.006) continue (0.006) future (0.006) which (0.006)
william	robert (0.003) james (0.003) john (0.003) richard (0.003) david (0.004) paul (0.005) michael (0.005) remains (0.005) recently (0.006) and (0.006)
workers	labor (0.007) jobs (0.008) employees (0.009) work (0.009) similar (0.010) union (0.010) local (0.010) changes (0.010) area (0.010) other (0.010)

Figure A.12 below shows the dendrogram containing the 1000 target words considered in analysis 12 of Chapter 4.

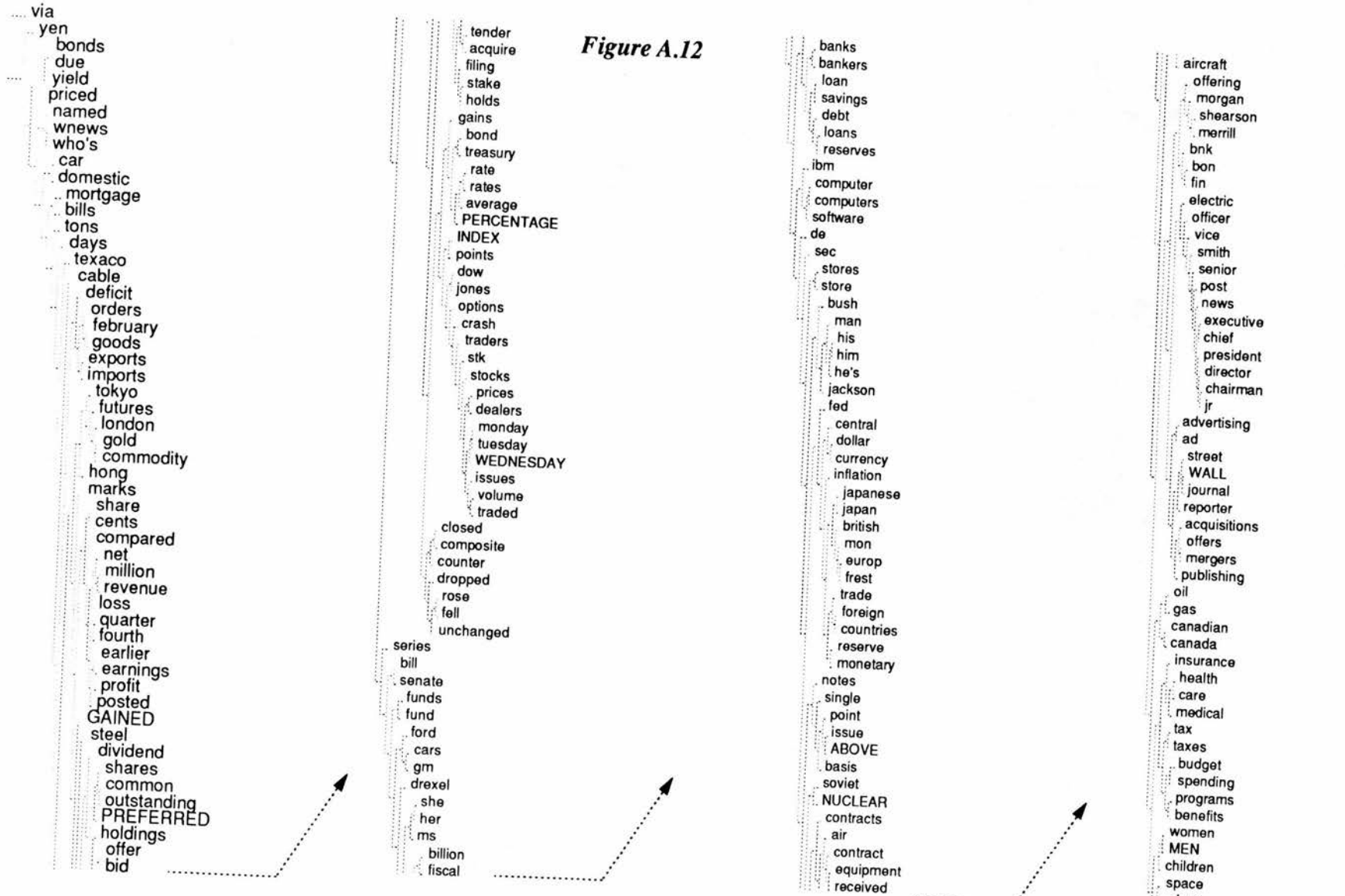

via
yen
bonds
due
yield
priced
named
wnews
who's
car
domestic
mortgage
bills
tons
days
texaco
cable
deficit
orders
february
goods
exports
imports
tokyo
futures
london
gold
commodity
hong
marks
share
cents
compared
net
million
revenue
loss
quarter
fourth
earlier
earnings
profit
posted
GAINED
steel
dividend
shares
common
outstanding
PREFERRED
holdings
offer
bid
tender
acquire
filing
stake
holds
gains
bond
treasury
rate
rates
average
PERCENTAGE
INDEX
points
dow
jones
options
crash
traders
stk
stocks
prices
dealers
monday
tuesday
WEDNESDAY
issues
volume
traded
closed
composite
counter
dropped
rose
fell
unchanged
series
bill
senate
funds
fund
ford
cars
gm
drexel
she
her
ms
billion
fiscal
banks
bankers
loan
savings
debt
loans
reserves
ibm
computer
computers
software
de
sec
stores
store
bush
man
his
him
he's
jackson
fed
central
dollar
currency
inflation
japanese
japan
british
mon
europ
frest
trade
foreign
countries
reserve
monetary
notes
single
point
issue
ABOVE
basis
soviet
NUCLEAR
contracts
air
contract
equipment
received
aircraft
offering
morgan
shearson
merrill
bnk
bon
fin
electric
officer
vice
smith
senior
post
news
executive
chief
president
director
chairman
jr
advertising
ad
street
WALL
journal
reporter
acquisitions
offers
mergers
publishing
oil
gas
canadian
canada
insurance
health
care
medical
tax
taxes
budget
spending
programs
benefits
women
MEN
children
space
class

Figure A.12

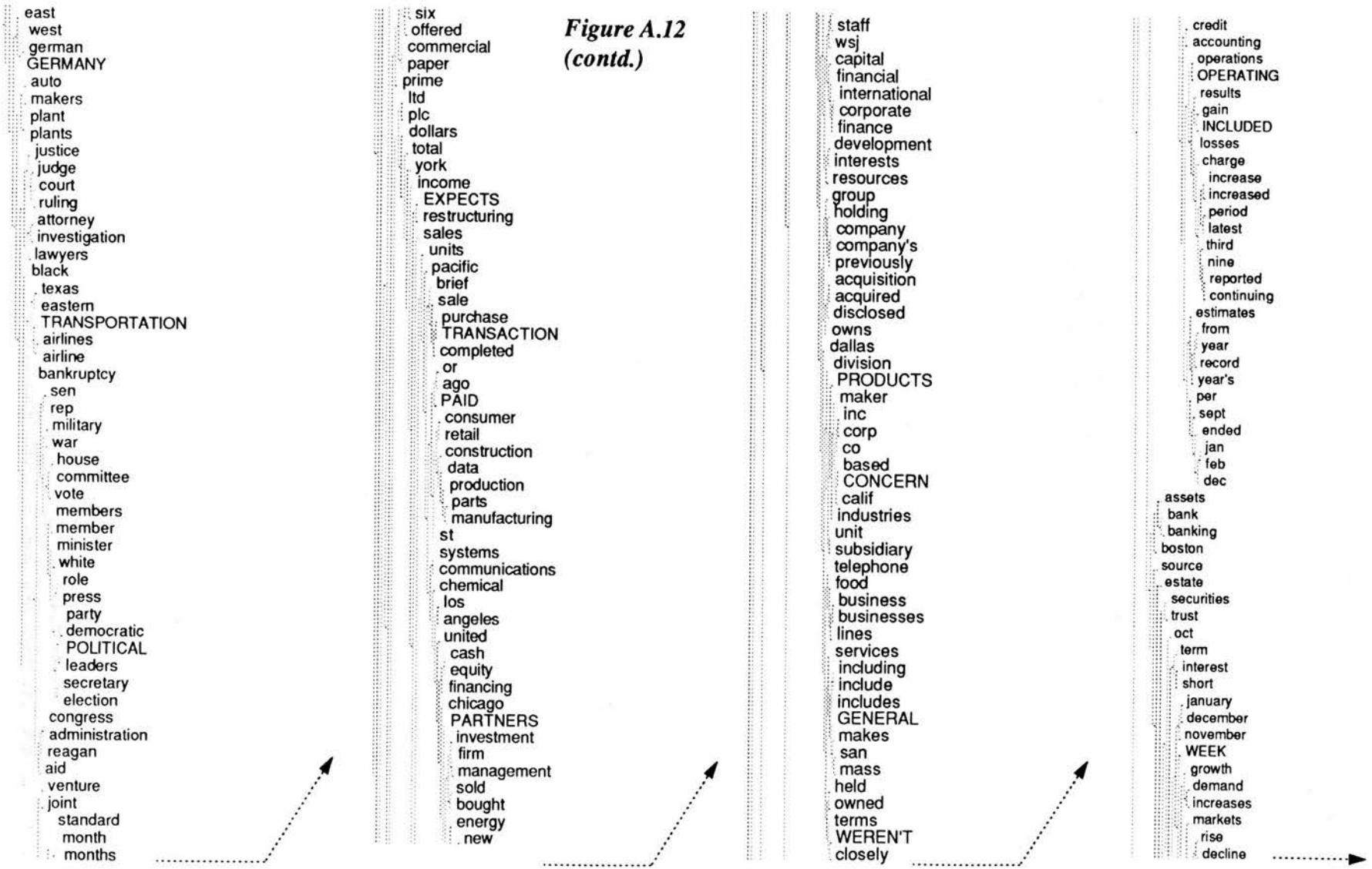


Figure A.12 (contd.)

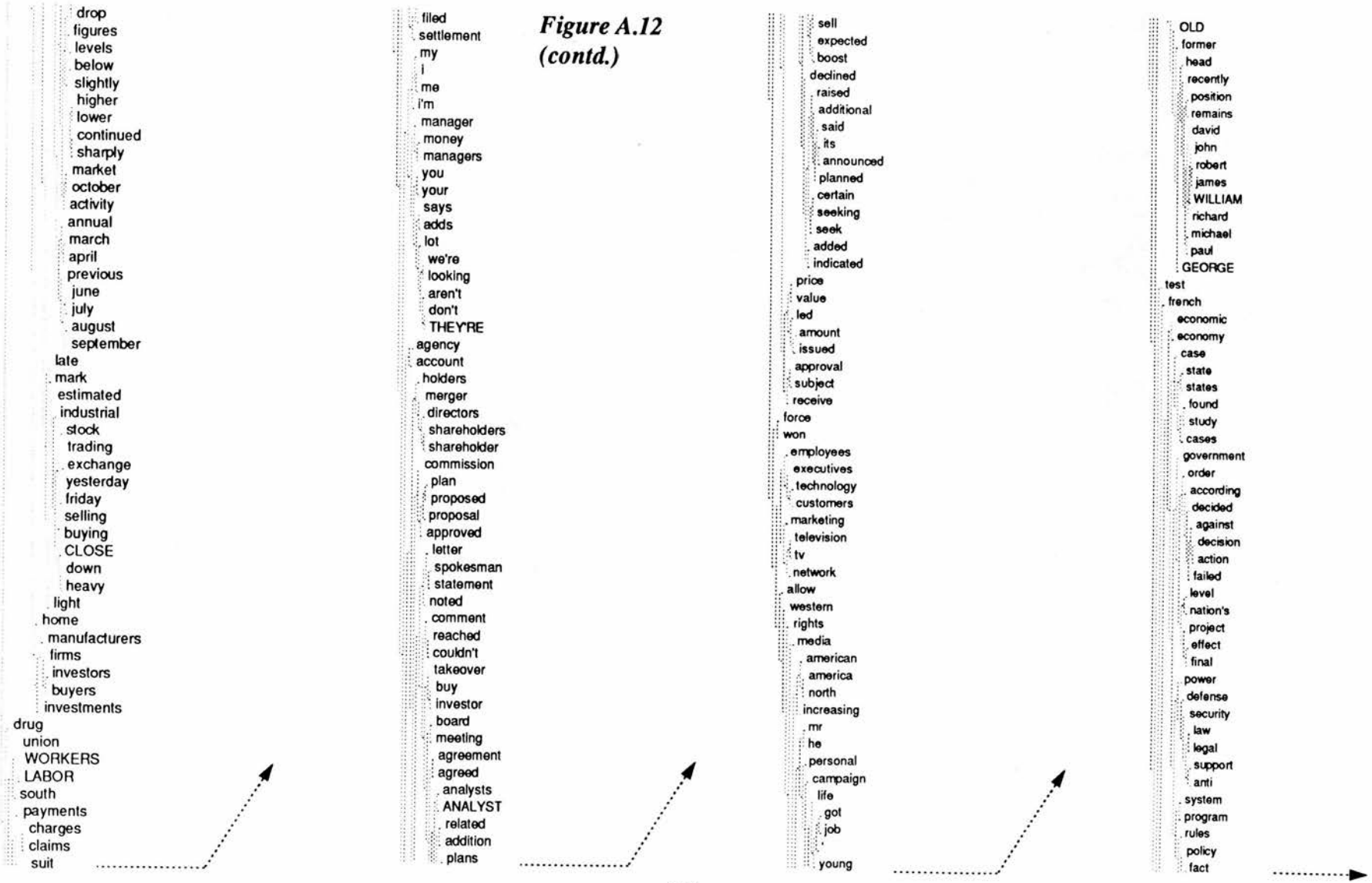


Figure A.12 (contd.)

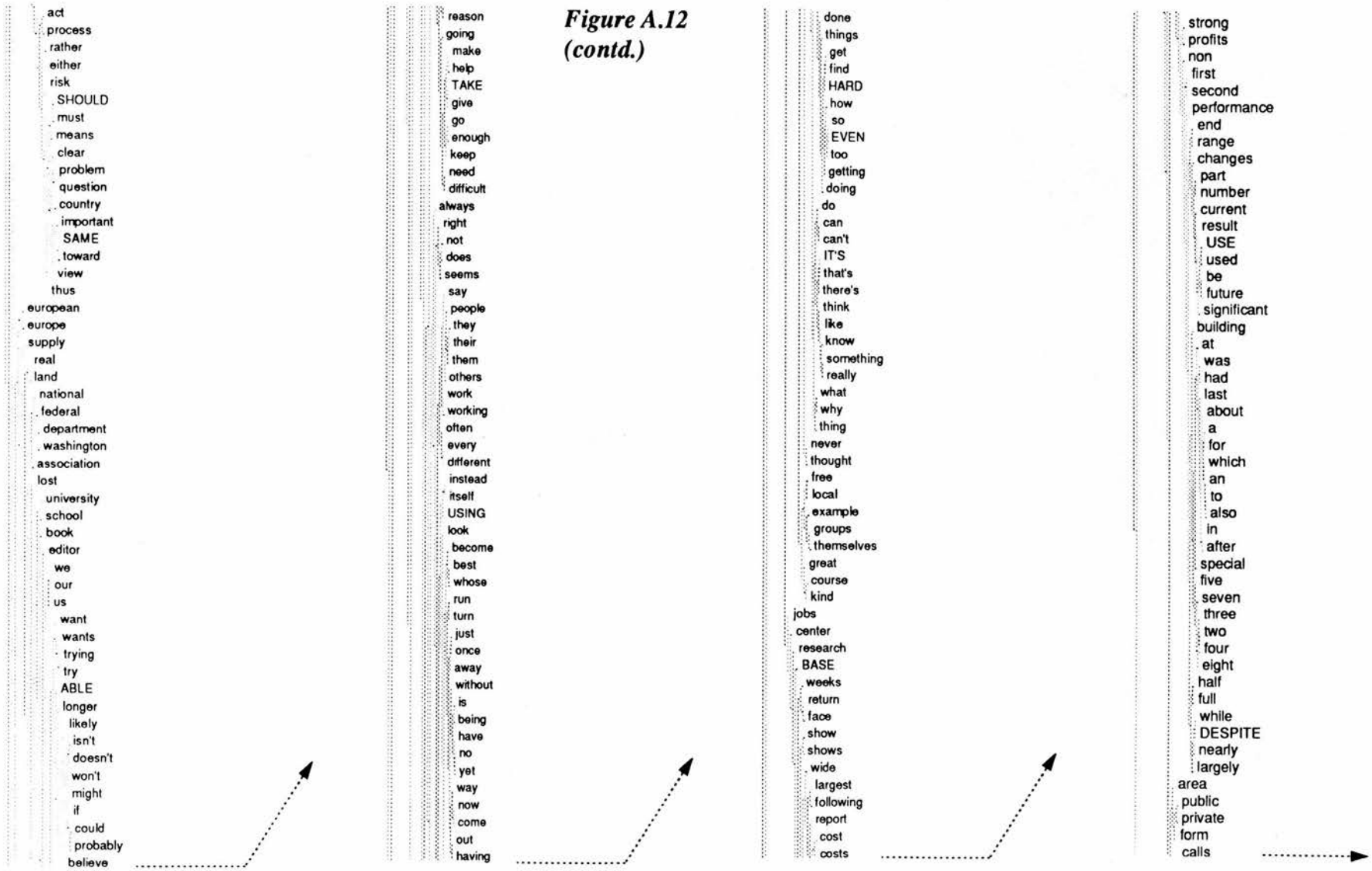


Figure A.12 (contd.)

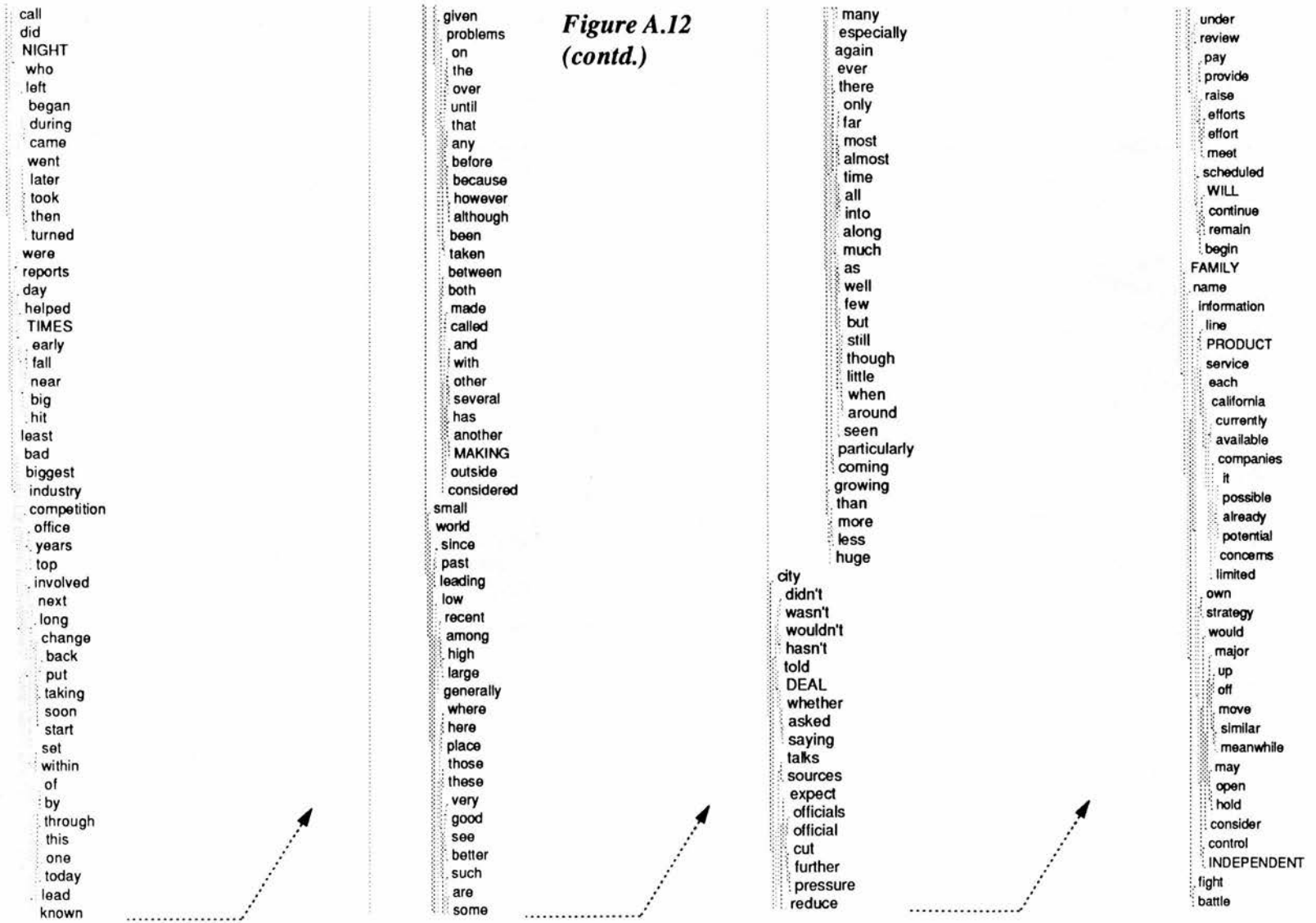


Figure A.12 (contd.)

APPENDIX B

TABLES OF NEAREST NEIGHBOURS

The tables which follow are tables of the nearest neighbours for each of the 29 lexical items considered in analyses 2-4, described in detail in chapter 7.

Table B.1

This table contains the nearest neighbours for the 29 lexical items from analysis 2.

Target Word	Nearest Neighbours (Euclidean Distance)
book	car (0.019) rock (0.028) girl (0.266) cat (0.267) boy (0.273) man (0.274) woman (0.278) mouse (0.285) dog (0.287) move (0.447) eat (0.451) exist (0.481) sleep (0.482) chase (0.490) think (0.491) like (0.496) monster (0.500) see (0.510) smell (0.512) lion (0.517) dragon (0.522) glass (0.535) plate (0.535) break (0.538) smash (0.538) bread (0.642) cookie (0.642) sandwich (0.642)
boy	girl (0.019) man (0.020) woman (0.028) mouse (0.111) dog (0.115) cat (0.117) rock (0.271) car (0.273) book (0.273) monster (0.298) lion (0.310) dragon (0.319) move (0.328) eat (0.332) sleep (0.401) exist (0.402) see (0.403) smell (0.403) chase (0.403) like (0.409) glass (0.410) plate (0.411) think (0.412) break (0.429) smash (0.429) sandwich (0.457) cookie (0.457) bread (0.457)
bread	sandwich (0.024) cookie (0.025) dragon (0.212) lion (0.218) monster (0.227) dog (0.421) mouse (0.430) cat (0.447) woman (0.448) man (0.456) girl (0.457) boy (0.457) move (0.538) eat (0.542) exist (0.557) sleep (0.557) think (0.565) chase (0.566) like (0.572) see (0.594) smell (0.594) plate (0.613) glass (0.613) break (0.621) smash (0.621) rock (0.639) car (0.641) book (0.642)
break	smash (0.023) eat (0.415) boy (0.429) exist (0.430) man (0.431) woman (0.432) sleep (0.434) girl (0.435) think (0.438) move (0.441) dog (0.474) mouse (0.475) cat (0.482) chase (0.482) see (0.492) smell (0.493) like (0.496) plate (0.511) glass (0.513) rock (0.537) book (0.538) car (0.541) monster (0.541) lion (0.546) dragon (0.557) bread (0.621) sandwich (0.621) cookie (0.622)
car	book (0.019) rock (0.027) girl (0.265) cat (0.267) boy (0.273) man (0.273) woman (0.277) mouse (0.285) dog (0.287) move (0.446) eat (0.449) exist (0.483) sleep (0.484) chase (0.487) like (0.492) think (0.494) monster (0.499) see (0.513) smell (0.514) lion (0.517) dragon (0.521) glass (0.534) plate (0.535) break (0.541) smash (0.541) cookie (0.641) sandwich (0.641) bread (0.641)
cat	mouse (0.031) dog (0.038) girl (0.111) boy (0.117) man (0.119) woman (0.121) rock (0.266) car (0.267) book (0.267) monster (0.303) lion (0.317) dragon (0.322) move (0.387) eat (0.387) sleep (0.445) exist (0.445) chase (0.446) sandwich (0.447) bread (0.447) cookie (0.447) like (0.451) think (0.455) see (0.457) smell (0.458) break (0.482) smash (0.482) glass (0.484) plate (0.485)
chase	like (0.059) sleep (0.215) exist (0.223) move (0.233) think (0.242) eat (0.336) see (0.390) smell (0.392) woman (0.395) man (0.400) girl (0.402) boy (0.403) glass (0.435) plate (0.440) mouse (0.442) dog (0.444) cat (0.446) rock (0.482) break (0.482) smash (0.483) car (0.487) book (0.490) monster (0.522) lion (0.527) dragon (0.541) cookie (0.562) sandwich (0.565) bread (0.566)
cookie	bread (0.025) sandwich (0.028) dragon (0.211) lion (0.218) monster (0.227) dog (0.421) mouse (0.430) cat (0.447) woman (0.447) man (0.456) girl (0.457) boy (0.457) move (0.537) eat (0.542) exist (0.557) sleep (0.557) chase (0.562) think (0.565) like (0.567) see (0.595) smell (0.596) glass (0.613) plate (0.613) break (0.622) smash (0.622) rock (0.639) car (0.641) book (0.642)
dog	mouse (0.022) cat (0.038) girl (0.111) boy (0.115) woman (0.118) man (0.118) monster (0.276) rock (0.286) book (0.287) car (0.287) lion (0.290) dragon (0.295) move (0.381) eat (0.387) sandwich (0.420) cookie (0.421) bread (0.421) sleep (0.441) exist (0.442) chase (0.444) like (0.449) smell (0.449) see (0.450) think (0.450) smash (0.474) break (0.474) glass (0.482) plate (0.483)
dragon	lion (0.031) monster (0.040) sandwich (0.210) cookie (0.211) bread (0.212) dog (0.295) mouse (0.305) woman (0.314) man (0.319) boy (0.319) girl (0.320) cat (0.322) move (0.480) eat (0.483) glass (0.516) plate (0.518) rock (0.521) car (0.521) book (0.522) see (0.532) smell (0.534) chase (0.541) sleep (0.543) like (0.544) exist (0.544) think (0.552) break (0.557) smash (0.557)
eat	move (0.293) woman (0.327) exist (0.329) girl (0.331) boy (0.332) man (0.332) sleep (0.332) chase (0.336) like (0.339) think (0.346) mouse (0.382) dog (0.387) cat (0.387) plate (0.410) glass (0.413) smash (0.415) break (0.415) see (0.425) smell (0.425) rock (0.447) car (0.449) book (0.451) lion (0.468) monster (0.470) dragon (0.483) sandwich (0.542) bread (0.542) cookie (0.542)
exist	sleep (0.036) think (0.074) chase (0.223) move (0.270) like (0.273) see (0.311) smell (0.314) eat (0.329) woman (0.395) man (0.399) girl (0.400) boy (0.402) break (0.430) plate (0.430) glass (0.430) smash (0.431) mouse (0.439) dog (0.442) cat (0.445) rock (0.477) book (0.481) car (0.483) monster (0.525) lion (0.530) dragon (0.544) bread (0.557) cookie (0.557) sandwich (0.559)

Table B.1 (contd.)

girl	man (0.018) boy (0.019) woman (0.025) mouse (0.106) dog (0.111) cat (0.111) rock (0.264) car (0.265) book (0.266) monster (0.298) lion (0.310) dragon (0.320) move (0.327) eat (0.331) sleep (0.399) exist (0.400) chase (0.402) see (0.402) smell (0.402) like (0.407) glass (0.408) plate (0.409) think (0.410) break (0.435) smash (0.435) sandwich (0.456) bread (0.457) cookie (0.457)
glass	plate (0.026) move (0.401) man (0.402) woman (0.403) girl (0.408) boy (0.410) eat (0.413) exist (0.430) sleep (0.431) chase (0.435) think (0.442) like (0.442) see (0.475) smell (0.477) mouse (0.479) dog (0.482) cat (0.484) monster (0.488) lion (0.496) break (0.513) smash (0.513) dragon (0.516) rock (0.532) car (0.534) book (0.535) cookie (0.613) bread (0.613) sandwich (0.613)
like	chase (0.059) move (0.239) sleep (0.265) exist (0.273) think (0.291) eat (0.339) woman (0.400) man (0.405) girl (0.407) boy (0.409) see (0.418) smell (0.419) glass (0.442) plate (0.447) mouse (0.447) dog (0.449) cat (0.451) rock (0.488) car (0.492) book (0.496) break (0.496) smash (0.497) monster (0.526) lion (0.531) dragon (0.544) cookie (0.567) sandwich (0.570) bread (0.572)
lion	monster (0.031) dragon (0.031) sandwich (0.217) cookie (0.218) bread (0.218) dog (0.290) mouse (0.300) woman (0.303) man (0.309) boy (0.310) girl (0.310) cat (0.317) move (0.467) eat (0.468) glass (0.496) plate (0.498) rock (0.515) car (0.517) book (0.517) smell (0.521) see (0.522) chase (0.527) sleep (0.529) exist (0.530) like (0.531) think (0.537) break (0.546) smash (0.546)
man	girl (0.018) boy (0.020) woman (0.023) mouse (0.113) dog (0.118) cat (0.119) rock (0.272) car (0.273) book (0.274) monster (0.297) lion (0.309) dragon (0.319) move (0.327) eat (0.332) sleep (0.398) exist (0.399) chase (0.400) glass (0.402) see (0.403) plate (0.404) smell (0.404) like (0.405) think (0.409) smash (0.431) break (0.431) sandwich (0.455) cookie (0.456) bread (0.456)
monster	lion (0.031) dragon (0.040) sandwich (0.226) cookie (0.227) bread (0.227) dog (0.276) mouse (0.286) woman (0.291) man (0.297) girl (0.298) boy (0.298) cat (0.303) move (0.461) eat (0.470) glass (0.488) plate (0.490) rock (0.498) car (0.499) book (0.500) see (0.516) smell (0.516) chase (0.522) sleep (0.524) exist (0.525) like (0.526) think (0.533) break (0.541) smash (0.541)
mouse	dog (0.022) cat (0.031) girl (0.106) boy (0.111) man (0.113) woman (0.114) rock (0.284) book (0.285) car (0.285) monster (0.286) lion (0.300) dragon (0.305) move (0.379) eat (0.382) sandwich (0.430) bread (0.430) cookie (0.430) sleep (0.438) exist (0.439) chase (0.442) see (0.447) smell (0.447) like (0.447) think (0.448) break (0.475) smash (0.475) glass (0.479) plate (0.480)
move	see (0.223) smell (0.225) chase (0.233) like (0.239) sleep (0.265) exist (0.270) think (0.284) eat (0.293) woman (0.323) man (0.327) girl (0.327) boy (0.328) mouse (0.379) dog (0.381) cat (0.387) glass (0.401) plate (0.405) break (0.441) smash (0.441) rock (0.443) car (0.446) book (0.447) monster (0.461) lion (0.467) dragon (0.480) cookie (0.537) sandwich (0.537) bread (0.538)
plate	glass (0.026) man (0.404) woman (0.404) move (0.405) girl (0.409) eat (0.410) boy (0.411) exist (0.430) sleep (0.431) chase (0.440) think (0.441) like (0.447) see (0.479) mouse (0.480) smell (0.480) dog (0.483) cat (0.485) monster (0.490) lion (0.498) break (0.511) smash (0.511) dragon (0.518) rock (0.532) car (0.535) book (0.535) bread (0.613) cookie (0.613) sandwich (0.613)
rock	car (0.027) book (0.028) girl (0.264) cat (0.266) boy (0.271) man (0.272) woman (0.274) mouse (0.284) dog (0.286) move (0.443) eat (0.447) sleep (0.477) exist (0.477) chase (0.482) think (0.485) like (0.488) monster (0.498) smell (0.510) see (0.511) lion (0.515) dragon (0.521) glass (0.532) plate (0.532) smash (0.537) break (0.537) sandwich (0.639) cookie (0.639) bread (0.639)
sandwich	bread (0.024) cookie (0.028) dragon (0.210) lion (0.217) monster (0.226) dog (0.420) mouse (0.430) cat (0.447) woman (0.448) man (0.455) girl (0.456) boy (0.457) move (0.537) eat (0.542) sleep (0.558) exist (0.559) chase (0.565) think (0.566) like (0.570) see (0.595) smell (0.596) glass (0.613) plate (0.613) break (0.621) smash (0.622) rock (0.639) car (0.641) book (0.642)
see	smell (0.063) move (0.223) sleep (0.309) exist (0.311) think (0.317) chase (0.390) girl (0.402) woman (0.402) boy (0.403) man (0.403) like (0.418) eat (0.425) mouse (0.447) dog (0.450) cat (0.457) glass (0.475) plate (0.479) break (0.492) smash (0.493) book (0.510) rock (0.511) car (0.513) monster (0.516) lion (0.522) dragon (0.532) bread (0.594) sandwich (0.595) cookie (0.595)
sleep	exist (0.036) think (0.065) chase (0.215) like (0.265) move (0.265) smell (0.308) see (0.309) eat (0.332) woman (0.394) man (0.398) girl (0.399) boy (0.401) glass (0.431) plate (0.431) break (0.434) smash (0.435) mouse (0.438) dog (0.441) cat (0.445) rock (0.477) book (0.482) car (0.484) monster (0.524) lion (0.529) dragon (0.543) bread (0.557) cookie (0.557) sandwich (0.558)
smash	break (0.023) eat (0.415) boy (0.429) man (0.431) exist (0.431) woman (0.432) sleep (0.435) girl (0.435) think (0.437) move (0.441) dog (0.474) mouse (0.475) cat (0.482) chase (0.483) see (0.493) smell (0.493) like (0.497) plate (0.511) glass (0.513) rock (0.537) book (0.538) car (0.541) monster (0.541) lion (0.546) dragon (0.557) bread (0.621) sandwich (0.622) cookie (0.622)
smell	see (0.063) move (0.225) sleep (0.308) think (0.309) exist (0.314) chase (0.392) woman (0.401) girl (0.402) boy (0.403) man (0.404) like (0.419) eat (0.425) mouse (0.447) dog (0.449) cat (0.458) glass (0.477) plate (0.480) break (0.493) smash (0.493) rock (0.510) book (0.512) car (0.514) monster (0.516) lion (0.521) dragon (0.534) bread (0.594) sandwich (0.596) cookie (0.596)
think	sleep (0.065) exist (0.074) chase (0.242) move (0.284) like (0.291) smell (0.309) see (0.317) eat (0.346) woman (0.405) man (0.409) girl (0.410) boy (0.412) smash (0.437) break (0.438) plate (0.441) glass (0.442) mouse (0.448) dog (0.450) cat (0.455) rock (0.485) book (0.491) car (0.494) monster (0.533) lion (0.537) dragon (0.552) bread (0.565) cookie (0.565) sandwich (0.566)
woman	man (0.023) girl (0.025) boy (0.028) mouse (0.114) dog (0.118) cat (0.121) rock (0.274) car (0.277) book (0.278) monster (0.291) lion (0.303) dragon (0.314) move (0.323) eat (0.327) sleep (0.394) chase (0.395) exist (0.395) like (0.400) smell (0.401) see (0.402) glass (0.403) plate (0.404) think (0.405) break (0.432) smash (0.432) cookie (0.447) sandwich (0.448) bread (0.448)

Table B.2

This table contains the nearest neighbours for the 29 lexical items from analysis 3.

Target Word	Nearest Neighbours (Spearman Correlation Coefficient)
book	rock (0.970) car (0.961) sandwich (0.896) cookie (0.877) bread (0.852) glass (0.778) plate (0.760) sleep (0.647) think (0.597) exist (0.570) man (0.448) woman (0.447) girl (0.444) cat (0.437) boy (0.407) move (0.404) chase (0.401) dog (0.400) mouse (0.387) like (0.380) see (0.272) smell (0.260) break (0.209) monster (0.203) smash (0.191) dragon (0.103) lion (0.078) eat (0.070)
boy	girl (0.962) man (0.957) woman (0.901) mouse (0.531) dog (0.516) cat (0.507) monster (0.465) lion (0.449) sleep (0.415) dragon (0.413) book (0.407) rock (0.399) car (0.392) move (0.354) think (0.348) glass (0.339) plate (0.299) sandwich (0.283) exist (0.276) bread (0.228) cookie (0.220) smell (0.169) see (0.168) chase (0.021) break (0.014) like (0.013) smash (-0.017) eat (-0.305)
bread	sandwich (0.954) cookie (0.950) book (0.852) rock (0.848) car (0.837) plate (0.833) glass (0.828) sleep (0.601) think (0.585) exist (0.584) cat (0.532) dog (0.465) mouse (0.427) move (0.417) chase (0.381) like (0.349) girl (0.315) woman (0.307) man (0.250) see (0.244) boy (0.228) smell (0.228) break (0.216) eat (0.216) smash (0.203) monster (0.072) dragon (-0.002) lion (-0.034)
break	smash (0.992) chase (0.796) like (0.785) see (0.729) smell (0.721) exist (0.683) eat (0.643) think (0.621) move (0.554) sleep (0.469) bread (0.216) book (0.209) sandwich (0.201) car (0.195) cookie (0.190) rock (0.167) monster (0.144) lion (0.091) cat (0.085) girl (0.081) dragon (0.070) glass (0.062) dog (0.059) woman (0.057) plate (0.037) man (0.036) mouse (0.027) boy (0.014)
car	rock (0.967) book (0.961) sandwich (0.888) cookie (0.880) bread (0.837) glass (0.771) plate (0.755) sleep (0.686) think (0.619) exist (0.580) cat (0.490) woman (0.473) move (0.445) man (0.444) chase (0.439) girl (0.429) dog (0.428) mouse (0.426) boy (0.392) like (0.390) see (0.286) smell (0.281) monster (0.228) break (0.195) smash (0.183) dragon (0.163) eat (0.119) lion (0.109)
cat	mouse (0.941) dog (0.939) move (0.698) sleep (0.563) bread (0.532) girl (0.531) woman (0.530) sandwich (0.519) cookie (0.513) boy (0.507) car (0.490) think (0.488) man (0.467) rock (0.459) book (0.437) exist (0.430) like (0.366) smell (0.345) see (0.344) chase (0.343) glass (0.324) plate (0.301) dragon (0.283) eat (0.254) monster (0.246) lion (0.196) smash (0.086) break (0.085)
chase	like (0.960) see (0.880) smell (0.877) break (0.796) move (0.793) smash (0.785) exist (0.755) eat (0.734) think (0.723) sleep (0.700) car (0.439) book (0.401) rock (0.395) cookie (0.392) sandwich (0.391) bread (0.381) cat (0.343) dog (0.331) mouse (0.312) glass (0.206) plate (0.167) dragon (0.135) monster (0.125) woman (0.123) lion (0.067) man (0.055) girl (0.053) boy (0.021)
cookie	bread (0.950) sandwich (0.942) car (0.880) book (0.877) rock (0.867) glass (0.841) plate (0.827) think (0.614) sleep (0.602) exist (0.590) cat (0.513) dog (0.433) mouse (0.421) move (0.392) chase (0.392) like (0.363) woman (0.340) girl (0.276) man (0.265) boy (0.220) see (0.218) smell (0.202) smash (0.192) break (0.190) eat (0.179) monster (0.151) dragon (0.041) lion (-0.009)
dog	mouse (0.963) cat (0.939) move (0.668) sleep (0.579) woman (0.550) girl (0.528) boy (0.516) sandwich (0.484) man (0.482) think (0.476) bread (0.465) exist (0.441) cookie (0.433) car (0.428) rock (0.423) book (0.400) like (0.374) dragon (0.352) smell (0.333) chase (0.331) lion (0.307) see (0.305) monster (0.299) eat (0.294) glass (0.272) plate (0.260) smash (0.061) break (0.059)
dragon	lion (0.944) monster (0.928) woman (0.442) sleep (0.438) man (0.437) boy (0.413) girl (0.380) mouse (0.368) exist (0.360) dog (0.352) think (0.319) move (0.304) cat (0.283) like (0.166) car (0.163) chase (0.135) rock (0.109) book (0.103) plate (0.079) glass (0.077) break (0.070) smash (0.054) smell (0.049) cookie (0.041) sandwich (0.030) see (0.025) eat (0.006) bread (-0.002)
eat	chase (0.734) like (0.724) smash (0.660) break (0.643) smell (0.638) see (0.628) move (0.551) sleep (0.516) exist (0.508) think (0.403) dog (0.294) cat (0.254) mouse (0.241) bread (0.216) sandwich (0.194) cookie (0.179) car (0.119) rock (0.075) book (0.070) dragon (0.006) glass (-0.016) plate (-0.025) lion (-0.026) monster (-0.083) woman (-0.207) girl (-0.253) boy (-0.305) man (-0.317)
exist	think (0.808) sleep (0.807) chase (0.755) like (0.736) move (0.702) break (0.683) smash (0.674) smell (0.664) see (0.629) sandwich (0.593) rock (0.590) cookie (0.590) bread (0.584) car (0.580) book (0.570) eat (0.508) plate (0.451) dog (0.441) glass (0.434) cat (0.430) monster (0.411) mouse (0.395) woman (0.375) girl (0.366) dragon (0.360) lion (0.336) man (0.324) boy (0.276)
girl	boy (0.962) man (0.950) woman (0.886) mouse (0.554) cat (0.531) dog (0.528) book (0.444) rock (0.439) sleep (0.437) car (0.429) monster (0.427) lion (0.420) glass (0.398) think (0.392) plate (0.380) dragon (0.380) move (0.376) exist (0.366) sandwich (0.332) bread (0.315) cookie (0.276) smell (0.205) see (0.190) break (0.081) like (0.057) smash (0.055) chase (0.053) eat (-0.253)
glass	plate (0.980) cookie (0.841) bread (0.828) sandwich (0.823) book (0.778) rock (0.775) car (0.771) think (0.479) sleep (0.466) exist (0.434) woman (0.406) girl (0.398) man (0.358) boy (0.339) cat (0.324) dog (0.272) mouse (0.247) move (0.221) chase (0.206) monster (0.196) like (0.151) see (0.080) dragon (0.077) lion (0.070) smash (0.067) smell (0.064) break (0.062) eat (-0.016)
like	chase (0.960) smell (0.883) see (0.866) move (0.817) break (0.785) smash (0.780) exist (0.736) think (0.728) eat (0.724) sleep (0.666) car (0.390) book (0.380) rock (0.376) dog (0.374) cat (0.366) cookie (0.363) mouse (0.362) bread (0.349) sandwich (0.349) monster (0.186) dragon (0.166) glass (0.151) lion (0.126) plate (0.125) woman (0.118) girl (0.057) man (0.042) boy (0.013)
lion	dragon (0.944) monster (0.943) man (0.476) woman (0.455) boy (0.449) sleep (0.428) girl (0.420) exist (0.336) mouse (0.317) dog (0.307) think (0.296) move (0.256) cat (0.196) like (0.126) car (0.109) break (0.091) rock (0.080) book (0.078) plate (0.073) smash (0.072) glass (0.070) chase (0.067) smell (0.033) see (0.005) cookie (-0.009) sandwich (-0.013) eat (-0.026) bread (-0.034)

Table B.2 (contd.)

man	boy (0.957) girl (0.950) woman (0.925) mouse (0.526) monster (0.506) dog (0.482) lion (0.476) cat (0.467) book (0.448) sleep (0.446) car (0.444) dragon (0.437) rock (0.415) glass (0.358) think (0.354) move (0.345) exist (0.324) plate (0.319) sandwich (0.305) cookie (0.265) bread (0.250) smell (0.186) see (0.185) chase (0.055) like (0.042) break (0.036) smash (0.005) eat (-0.317)
monster	lion (0.943) dragon (0.928) woman (0.508) man (0.506) boy (0.465) sleep (0.441) girl (0.427) exist (0.411) think (0.381) mouse (0.318) dog (0.299) move (0.285) cat (0.246) car (0.228) book (0.203) glass (0.196) rock (0.196) plate (0.190) like (0.186) cookie (0.151) break (0.144) smash (0.127) chase (0.125) sandwich (0.102) bread (0.072) smell (0.056) see (0.032) eat (-0.083)
mouse	dog (0.963) cat (0.941) move (0.670) woman (0.571) sleep (0.567) girl (0.554) boy (0.531) man (0.526) think (0.469) bread (0.427) car (0.426) cookie (0.421) sandwich (0.421) exist (0.395) rock (0.395) book (0.387) dragon (0.368) like (0.362) smell (0.339) see (0.320) monster (0.318) lion (0.317) chase (0.312) glass (0.247) eat (0.241) plate (0.230) smash (0.029) break (0.027)
move	see (0.859) smell (0.857) like (0.817) chase (0.793) sleep (0.740) exist (0.702) cat (0.698) think (0.697) mouse (0.670) dog (0.668) break (0.554) eat (0.551) smash (0.543) car (0.445) bread (0.417) sandwich (0.406) rock (0.405) book (0.404) cookie (0.392) woman (0.389) girl (0.376) boy (0.354) man (0.345) dragon (0.304) monster (0.285) lion (0.256) glass (0.221) plate (0.185)
plate	glass (0.980) bread (0.833) cookie (0.827) sandwich (0.827) rock (0.780) book (0.760) car (0.755) think (0.468) exist (0.451) sleep (0.437) girl (0.380) woman (0.377) man (0.319) cat (0.301) boy (0.299) dog (0.260) mouse (0.230) monster (0.190) move (0.185) chase (0.167) like (0.125) dragon (0.079) lion (0.073) smash (0.040) break (0.037) smell (0.033) see (0.029) eat (-0.025)
rock	book (0.970) car (0.967) sandwich (0.895) cookie (0.867) bread (0.848) plate (0.780) glass (0.775) sleep (0.653) think (0.617) exist (0.590) cat (0.459) woman (0.455) girl (0.439) dog (0.423) man (0.415) move (0.405) boy (0.399) chase (0.395) mouse (0.395) like (0.376) smell (0.259) see (0.248) monster (0.196) break (0.167) smash (0.150) dragon (0.109) lion (0.080) eat (0.075)
sandwich	bread (0.954) cookie (0.942) book (0.896) rock (0.895) car (0.888) plate (0.827) glass (0.823) sleep (0.626) exist (0.593) think (0.588) cat (0.519) dog (0.484) mouse (0.421) move (0.406) chase (0.391) woman (0.349) like (0.349) girl (0.332) man (0.305) boy (0.283) see (0.229) smell (0.219) break (0.201) eat (0.194) smash (0.189) monster (0.102) dragon (0.030) lion (-0.013)
see	smell (0.977) chase (0.880) like (0.866) move (0.859) break (0.729) smash (0.708) sleep (0.639) exist (0.629) eat (0.628) think (0.595) cat (0.344) mouse (0.320) dog (0.305) car (0.286) book (0.272) rock (0.248) bread (0.244) sandwich (0.229) cookie (0.218) woman (0.210) girl (0.190) man (0.185) boy (0.168) glass (0.080) monster (0.032) plate (0.029) dragon (0.025) lion (0.005)
sleep	exist (0.807) think (0.766) move (0.740) chase (0.700) car (0.686) like (0.666) rock (0.653) book (0.647) see (0.639) smell (0.633) sandwich (0.626) cookie (0.602) bread (0.601) dog (0.579) mouse (0.567) cat (0.563) eat (0.516) woman (0.479) break (0.469) glass (0.466) man (0.446) monster (0.441) dragon (0.438) girl (0.437) plate (0.437) smash (0.433) lion (0.428) boy (0.415)
smash	break (0.992) chase (0.785) like (0.780) see (0.708) smell (0.703) exist (0.674) eat (0.660) think (0.615) move (0.543) sleep (0.433) bread (0.203) cookie (0.192) book (0.191) sandwich (0.189) car (0.183) rock (0.150) monster (0.127) cat (0.086) lion (0.072) glass (0.067) dog (0.061) girl (0.055) dragon (0.054) plate (0.040) woman (0.039) mouse (0.029) man (0.005) boy (-0.017)
smell	see (0.977) like (0.883) chase (0.877) move (0.857) break (0.721) smash (0.703) exist (0.664) eat (0.638) sleep (0.633) think (0.615) cat (0.345) mouse (0.339) dog (0.333) car (0.281) book (0.260) rock (0.259) woman (0.239) bread (0.228) sandwich (0.219) girl (0.205) cookie (0.202) man (0.186) boy (0.169) glass (0.064) monster (0.056) dragon (0.049) lion (0.033) plate (0.033)
think	exist (0.808) sleep (0.766) like (0.728) chase (0.723) move (0.697) break (0.621) car (0.619) rock (0.617) smash (0.615) smell (0.615) cookie (0.614) book (0.597) see (0.595) sandwich (0.588) bread (0.585) cat (0.488) glass (0.479) woman (0.476) dog (0.476) mouse (0.469) plate (0.468) eat (0.403) girl (0.392) monster (0.381) man (0.354) boy (0.348) dragon (0.319) lion (0.296)
woman	man (0.925) boy (0.901) girl (0.886) mouse (0.571) dog (0.550) cat (0.530) monster (0.508) sleep (0.479) think (0.476) car (0.473) rock (0.455) lion (0.455) book (0.447) dragon (0.442) glass (0.406) move (0.389) plate (0.377) exist (0.375) sandwich (0.349) cookie (0.340) bread (0.307) smell (0.239) see (0.210) chase (0.123) like (0.118) break (0.057) smash (0.039) eat (-0.207)

Table B.3

This table contains the nearest neighbours for the 29 lexical items from analysis 4.

Target Word	Nearest Neighbours (Euclidean Distance)
book	rock (0.020) car (0.021) sleep (0.152) girl (0.162) cat (0.165) exist (0.166) man (0.167) boy (0.168) woman (0.169) think (0.171) mouse (0.171) dog (0.172) move (0.199) glass (0.256) plate (0.264) chase (0.266) eat (0.272) like (0.274) monster (0.275) smell (0.293) dragon (0.293) lion (0.295) see (0.296) bread (0.306) break (0.307) smash (0.308) cookie (0.310) sandwich (0.310)

Table B.3 (contd.)

boy	man (0.013) girl (0.015) woman (0.021) mouse (0.085) dog (0.085) cat (0.088) exist (0.149) sleep (0.151) think (0.151) move (0.155) monster (0.157) book (0.168) car (0.174) rock (0.175) lion (0.175) dragon (0.179) glass (0.215) plate (0.216) bread (0.232) cookie (0.235) eat (0.235) sandwich (0.238) smell (0.248) see (0.249) break (0.260) smash (0.262) chase (0.271) like (0.272)
bread	cookie (0.018) sandwich (0.019) dragon (0.143) lion (0.152) monster (0.153) dog (0.208) mouse (0.211) cat (0.216) woman (0.223) girl (0.227) boy (0.232) man (0.233) think (0.272) exist (0.279) sleep (0.280) car (0.306) book (0.306) plate (0.307) move (0.309) rock (0.313) glass (0.317) eat (0.356) like (0.371) chase (0.371) break (0.390) smash (0.391) smell (0.392) see (0.393)
break	smash (0.024) eat (0.207) exist (0.221) move (0.223) think (0.230) sleep (0.236) like (0.238) chase (0.239) see (0.246) smell (0.246) man (0.258) boy (0.260) woman (0.261) girl (0.264) glass (0.296) plate (0.299) dog (0.304) mouse (0.304) cat (0.305) book (0.307) rock (0.311) car (0.315) monster (0.329) lion (0.344) dragon (0.350) bread (0.390) cookie (0.392) sandwich (0.396)
car	book (0.021) rock (0.025) sleep (0.161) cat (0.166) girl (0.168) mouse (0.172) dog (0.173) man (0.174) boy (0.174) exist (0.174) woman (0.175) think (0.180) move (0.205) glass (0.266) chase (0.272) plate (0.274) monster (0.277) eat (0.278) like (0.280) dragon (0.294) lion (0.296) smell (0.301) see (0.304) bread (0.306) cookie (0.309) sandwich (0.310) break (0.315) smash (0.315)
cat	mouse (0.019) dog (0.022) girl (0.082) woman (0.086) boy (0.088) man (0.091) book (0.165) car (0.166) monster (0.169) rock (0.171) move (0.180) lion (0.184) dragon (0.186) sleep (0.186) exist (0.192) think (0.192) bread (0.216) cookie (0.219) sandwich (0.221) glass (0.253) plate (0.254) eat (0.255) like (0.284) chase (0.287) smell (0.294) see (0.296) break (0.305) smash (0.307)
chase	like (0.043) move (0.157) eat (0.172) sleep (0.173) exist (0.177) think (0.183) smell (0.198) see (0.202) break (0.239) smash (0.242) book (0.266) man (0.268) rock (0.268) woman (0.269) boy (0.271) car (0.272) girl (0.273) glass (0.285) cat (0.287) mouse (0.288) dog (0.290) plate (0.290) monster (0.337) lion (0.351) dragon (0.353) cookie (0.370) bread (0.371) sandwich (0.375)
cookie	bread (0.018) sandwich (0.023) dragon (0.142) lion (0.151) monster (0.152) dog (0.211) mouse (0.214) cat (0.219) woman (0.226) girl (0.230) boy (0.235) man (0.235) think (0.273) exist (0.280) sleep (0.282) plate (0.309) car (0.309) move (0.310) book (0.310) rock (0.316) glass (0.319) eat (0.357) like (0.368) chase (0.370) break (0.392) smell (0.392) smash (0.392) see (0.393)
dog	cat (0.022) mouse (0.025) girl (0.080) woman (0.083) boy (0.085) man (0.089) monster (0.157) lion (0.172) book (0.172) car (0.173) dragon (0.174) rock (0.177) move (0.184) sleep (0.187) exist (0.193) think (0.193) bread (0.208) cookie (0.211) sandwich (0.213) glass (0.252) plate (0.253) eat (0.256) like (0.288) chase (0.290) smell (0.297) see (0.299) break (0.304) smash (0.306)
dragon	lion (0.024) monster (0.033) cookie (0.142) bread (0.143) sandwich (0.145) woman (0.174) dog (0.174) mouse (0.177) girl (0.177) boy (0.179) man (0.180) cat (0.186) think (0.252) exist (0.256) sleep (0.259) move (0.272) plate (0.290) book (0.293) car (0.294) glass (0.298) rock (0.299) eat (0.330) break (0.350) like (0.351) smash (0.351) chase (0.353) smell (0.358) see (0.359)
eat	move (0.163) like (0.168) chase (0.172) exist (0.191) sleep (0.192) think (0.201) break (0.207) smash (0.209) smell (0.212) see (0.213) woman (0.234) man (0.234) boy (0.235) girl (0.238) mouse (0.254) cat (0.255) dog (0.256) glass (0.258) plate (0.261) book (0.272) rock (0.275) car (0.278) monster (0.316) lion (0.325) dragon (0.330) bread (0.356) cookie (0.357) sandwich (0.362)
exist	sleep (0.034) think (0.038) move (0.137) man (0.144) woman (0.145) girl (0.148) boy (0.149) book (0.166) rock (0.170) car (0.174) chase (0.177) smell (0.189) eat (0.191) cat (0.192) like (0.192) mouse (0.192) dog (0.193) see (0.193) glass (0.212) plate (0.214) break (0.221) smash (0.222) monster (0.238) lion (0.256) dragon (0.256) bread (0.279) cookie (0.280) sandwich (0.284)
girl	boy (0.015) man (0.017) woman (0.021) mouse (0.079) dog (0.080) cat (0.082) exist (0.148) sleep (0.150) think (0.151) monster (0.156) move (0.157) book (0.162) car (0.168) rock (0.169) lion (0.174) dragon (0.177) glass (0.216) plate (0.217) bread (0.227) cookie (0.230) sandwich (0.233) eat (0.238) smell (0.251) see (0.253) break (0.264) smash (0.266) chase (0.273) like (0.274)
glass	plate (0.027) man (0.211) exist (0.212) woman (0.213) boy (0.215) girl (0.216) sleep (0.217) think (0.217) move (0.239) mouse (0.251) dog (0.252) cat (0.253) book (0.256) eat (0.258) rock (0.262) car (0.266) monster (0.277) chase (0.285) like (0.289) lion (0.293) see (0.294) smell (0.295) smash (0.295) break (0.296) dragon (0.298) bread (0.317) cookie (0.319) sandwich (0.326)
like	chase (0.043) move (0.154) eat (0.168) sleep (0.187) exist (0.192) think (0.196) smell (0.203) see (0.209) break (0.238) smash (0.242) man (0.269) woman (0.270) boy (0.272) book (0.274) girl (0.274) rock (0.275) car (0.280) cat (0.284) mouse (0.285) dog (0.288) glass (0.289) plate (0.294) monster (0.335) lion (0.349) dragon (0.351) cookie (0.368) bread (0.371) sandwich (0.375)
lion	dragon (0.024) monster (0.029) cookie (0.151) bread (0.152) sandwich (0.155) woman (0.170) dog (0.172) girl (0.174) mouse (0.175) boy (0.175) man (0.176) cat (0.184) think (0.251) exist (0.256) sleep (0.258) move (0.270) plate (0.285) glass (0.293) book (0.295) car (0.296) rock (0.300) eat (0.325) break (0.344) smash (0.346) like (0.349) chase (0.351) smell (0.356) see (0.358)
man	boy (0.013) girl (0.017) woman (0.020) mouse (0.087) dog (0.089) cat (0.091) exist (0.144) sleep (0.147) think (0.147) move (0.154) monster (0.157) book (0.167) car (0.174) rock (0.174) lion (0.176) dragon (0.180) glass (0.211) plate (0.212) bread (0.233) eat (0.234) cookie (0.235) sandwich (0.238) smell (0.246) see (0.248) break (0.258) smash (0.259) chase (0.268) like (0.269)
monster	lion (0.029) dragon (0.033) cookie (0.152) woman (0.152) bread (0.153) girl (0.156) sandwich (0.156) dog (0.157) boy (0.157) man (0.157) mouse (0.160) cat (0.169) think (0.234) exist (0.238) sleep (0.241) move (0.253) plate (0.270) book (0.275) car (0.277) glass (0.277) rock (0.281) eat (0.316) break (0.329) smash (0.331) like (0.335) chase (0.337) smell (0.341) see (0.343)
mouse	cat (0.019) dog (0.025) girl (0.079) woman (0.082) boy (0.085) man (0.087) monster (0.160) book (0.171) car (0.172) lion (0.175) rock (0.177) dragon (0.177) move (0.180) sleep (0.187) exist (0.192) think (0.192) bread (0.211) cookie (0.214) sandwich (0.216) glass (0.251) plate (0.252) eat (0.254) like (0.285) chase (0.288) smell (0.294) see (0.296) break (0.304) smash (0.306)

Table B.3 (contd.)

move	sleep (0.134) smell (0.136) exist (0.137) see (0.139) think (0.144) man (0.154) like (0.154) boy (0.155) woman (0.157) girl (0.157) chase (0.157) eat (0.163) cat (0.180) mouse (0.180) dog (0.184) book (0.199) rock (0.203) car (0.205) break (0.223) smash (0.226) glass (0.239) plate (0.245) monster (0.253) lion (0.270) dragon (0.272) bread (0.309) cookie (0.310) sandwich (0.314)
plate	glass (0.027) man (0.212) woman (0.213) exist (0.214) boy (0.216) girl (0.217) think (0.217) sleep (0.220) move (0.245) mouse (0.252) dog (0.253) cat (0.254) eat (0.261) book (0.264) rock (0.270) monster (0.270) car (0.274) lion (0.285) dragon (0.290) chase (0.290) like (0.294) smash (0.298) break (0.299) smell (0.299) see (0.299) bread (0.307) cookie (0.309) sandwich (0.316)
rock	book (0.020) car (0.025) sleep (0.156) girl (0.169) exist (0.170) cat (0.171) think (0.174) man (0.174) boy (0.175) woman (0.175) mouse (0.177) dog (0.177) move (0.203) glass (0.262) chase (0.268) plate (0.270) like (0.275) eat (0.275) monster (0.281) smell (0.296) dragon (0.299) lion (0.300) see (0.301) break (0.311) smash (0.312) bread (0.313) cookie (0.316) sandwich (0.316)
sandwich	bread (0.019) cookie (0.023) dragon (0.145) lion (0.155) monster (0.156) dog (0.213) mouse (0.216) cat (0.221) woman (0.229) girl (0.233) boy (0.238) man (0.238) think (0.276) exist (0.284) sleep (0.285) car (0.310) book (0.310) move (0.314) rock (0.316) plate (0.316) glass (0.326) eat (0.362) like (0.375) chase (0.375) break (0.396) smell (0.396) smash (0.397) see (0.397)
see	smell (0.040) move (0.139) exist (0.193) sleep (0.199) chase (0.202) think (0.202) like (0.209) eat (0.213) break (0.246) man (0.248) smash (0.248) boy (0.249) woman (0.252) girl (0.253) glass (0.294) mouse (0.296) cat (0.296) book (0.296) dog (0.299) plate (0.299) rock (0.301) car (0.304) monster (0.343) lion (0.358) dragon (0.359) bread (0.393) cookie (0.393) sandwich (0.397)
sleep	exist (0.034) think (0.040) move (0.134) man (0.147) woman (0.147) girl (0.150) boy (0.151) book (0.152) rock (0.156) car (0.161) chase (0.173) cat (0.186) mouse (0.187) dog (0.187) like (0.187) eat (0.192) smell (0.193) see (0.199) glass (0.217) plate (0.220) break (0.236) smash (0.237) monster (0.241) lion (0.258) dragon (0.259) bread (0.280) cookie (0.282) sandwich (0.285)
smash	break (0.024) eat (0.209) exist (0.222) move (0.226) think (0.231) sleep (0.237) like (0.242) chase (0.242) see (0.248) smell (0.249) man (0.259) boy (0.262) woman (0.262) girl (0.266) glass (0.295) plate (0.298) mouse (0.306) dog (0.306) cat (0.307) book (0.308) rock (0.312) car (0.315) monster (0.331) lion (0.346) dragon (0.351) bread (0.391) cookie (0.392) sandwich (0.397)
smell	see (0.040) move (0.136) exist (0.189) sleep (0.193) think (0.195) chase (0.198) like (0.203) eat (0.212) man (0.246) break (0.246) boy (0.248) smash (0.249) woman (0.250) girl (0.251) book (0.293) mouse (0.294) cat (0.294) glass (0.295) rock (0.296) dog (0.297) plate (0.299) car (0.301) monster (0.341) lion (0.356) dragon (0.358) bread (0.392) cookie (0.392) sandwich (0.396)
think	exist (0.038) sleep (0.040) move (0.144) woman (0.145) man (0.147) girl (0.151) boy (0.151) book (0.171) rock (0.174) car (0.180) chase (0.183) mouse (0.192) cat (0.192) dog (0.193) smell (0.195) like (0.196) eat (0.201) see (0.202) glass (0.217) plate (0.217) break (0.230) smash (0.231) monster (0.234) lion (0.251) dragon (0.252) bread (0.272) cookie (0.273) sandwich (0.276)
woman	man (0.020) girl (0.021) boy (0.021) mouse (0.082) dog (0.083) cat (0.086) exist (0.145) think (0.145) sleep (0.147) monster (0.152) move (0.157) book (0.169) lion (0.170) dragon (0.174) car (0.175) rock (0.175) glass (0.213) plate (0.213) bread (0.223) cookie (0.226) sandwich (0.229) eat (0.234) smell (0.250) see (0.252) break (0.261) smash (0.262) chase (0.269) like (0.270)

APPENDIX C: DENDROGRAM

This appendix contains the table of randomly selected target words and the dendrogram for the 1000 target words analysed by the unsupervised neural network described in Chapter 8, based on their distributional similarity over the output units of the network (using the Spearman Rank Correlation Coefficient as the distance metric).

The table of nearest neighbours is illustrated in table C.1 below

Table C.1: Nearest Neighbours for Randomly Selected Target Words Considered in Chapter 8

Target Word	10 Nearest Neighbours (Spearman Correlation Coefficient)
able	rights (0.828) better (0.809) plan (0.799) telephone (0.799) estimated (0.793) tender (0.791) tax (0.785) dividend (0.782) yield (0.781) maker (0.779)
above	really (0.894) among (0.880) indicated (0.845) noted (0.844) pay (0.842) wasn't (0.841) least (0.828) levels (0.827) face (0.824) calif (0.817)
analyst	she (0.893) administration (0.876) commission (0.872) systems (0.872) without (0.867) compared (0.863) by (0.861) editor (0.858) issue (0.856) problem (0.856)
close	yield (0.940) rise (0.928) sell (0.926) charge (0.922) cut (0.916) telephone (0.915) 4 (0.913) move (0.911) bid (0.908) start (0.903)
concern	rules (0.962) business (0.949) campaign (0.947) claims (0.944) line (0.943) today (0.941) system (0.940) sale (0.940) journal (0.939) jobs (0.936)
deal	bid (0.944) report (0.942) range (0.939) post (0.927) share (0.924) review (0.919) yield (0.916) position (0.911) study (0.911) question (0.910)
despite	within (0.874) under (0.848) because (0.841) following (0.840) after (0.835) if (0.823) before (0.821) without (0.819) include (0.814) announced (0.809)
even	so (0.926) friday (0.903) closed (0.899) not (0.897) recently (0.895) which (0.893) spending (0.890) is (0.889) groups (0.888) costs (0.885)
expects	fell (0.932) disclosed (0.928) done (0.921) wants (0.918) called (0.916) sold (0.909) got (0.905) also (0.894) were (0.892) been (0.891)
family	performance (0.984) unit (0.979) drug (0.977) record (0.966) low (0.964) development (0.950) university (0.948) california (0.946) service (0.944) credit (0.939)
gained	made (0.889) remain (0.881) were (0.877) bought (0.865) isn't (0.860) it's (0.859) acquire (0.857) took (0.853) still (0.842) led (0.838)
general	commercial (0.970) national (0.964) higher (0.958) big (0.954) future (0.953) one (0.942) cash (0.941) standard (0.939) current (0.936) heavy (0.932)
germany	company (0.925) problem (0.914) day (0.913) plant (0.911) officer (0.907) loss (0.906) meeting (0.905) system (0.902) party (0.902) director (0.901)
hard	plan (0.935) company (0.925) common (0.919) world (0.916) management (0.915) loss (0.915) possible (0.915) strategy (0.913) firm (0.913) industries (0.912)
included	isn't (0.919) took (0.915) currently (0.908) it's (0.894) includes (0.887) began (0.881) there's (0.877) received (0.866) provide (0.862) would (0.861)
index	administration (0.944) problem (0.933) director (0.929) plant (0.926) bill (0.926) air (0.925) department (0.925) value (0.924) price (0.924) company (0.922)
it's	isn't (0.944) there's (0.915) included (0.894) took (0.893) that's (0.889) is (0.879) weren't (0.876) currently (0.876) probably (0.868) are (0.867)
labor	national (0.883) health (0.879) state (0.878) consumer (0.875) futures (0.873) software (0.872) political (0.872) future (0.871) standard (0.871) gold (0.866)

Table C.1 (contd.)

making	getting (0.950) using (0.930) selling (0.921) taking (0.904) offering (0.897) being (0.896) increased (0.892) then (0.890) too (0.885) april (0.885)
nuclear	former (0.959) corporate (0.958) fiscal (0.958) longterm (0.939) personal (0.929) democratic (0.923) previous (0.918) north (0.900) current (0.900) jan (0.892)
old	next (0.928) growth (0.924) stocks (0.922) capital (0.921) software (0.911) equipment (0.906) world (0.904) television (0.903) equity (0.903) common (0.901)
operating	working (0.939) exchange (0.932) spending (0.930) industries (0.929) right (0.928) production (0.927) employees (0.925) investor (0.921) building (0.917) orders (0.916)
paid	issued (0.845) offered (0.825) disclosed (0.809) approved (0.808) lost (0.807) decision (0.795) rose (0.792) still (0.792) left (0.791) over (0.789)
partners	friday (0.962) rules (0.955) jobs (0.952) crash (0.952) terms (0.951) acquisition (0.949) units (0.946) system (0.946) changes (0.945) today (0.943)
percentage	buyout (0.912) filing (0.882) case (0.878) basis (0.876) war (0.875) venture (0.870) merger (0.866) letter (0.860) banking (0.860) fact (0.850)
political	software (0.955) capital (0.951) legal (0.950) media (0.950) state (0.949) black (0.949) national (0.947) television (0.945) big (0.942) common (0.940)
preferred	dividend (0.963) security (0.946) product (0.931) software (0.923) securities (0.921) chemical (0.920) revenue (0.917) bank (0.916) planned (0.913) capital (0.911)
product	business (0.977) security (0.975) revenue (0.972) loss (0.970) market (0.968) party (0.966) service (0.964) job (0.964) home (0.961) system (0.958)
products	goods (0.951) shares (0.950) costs (0.947) problems (0.937) industries (0.937) leaders (0.932) assets (0.931) concerns (0.931) stocks (0.928) outstanding (0.925)
same	industrial (0.818) e (0.815) r (0.808) company's (0.801) economic (0.798) ibm (0.797) overthecounter (0.791) latest (0.777) jan (0.769) vice (0.763)
should	could (0.961) didn't (0.960) makes (0.949) can (0.938) can't (0.919) are (0.912) wouldn't (0.910) until (0.908) would (0.906) was (0.901)
take	find (0.953) do (0.953) have (0.950) acquire (0.933) know (0.930) see (0.928) give (0.924) took (0.923) be (0.917) remain (0.905)
times	performance (0.939) family (0.931) acquired (0.928) states (0.925) unit (0.925) monday (0.920) drug (0.919) february (0.917) december (0.911) face (0.909)
transaction	statement (0.951) strategy (0.946) number (0.940) meeting (0.931) day (0.930) group (0.928) way (0.927) city (0.926) ruling (0.923) system (0.918)
use	turn (0.941) others (0.920) gain (0.916) average (0.916) review (0.915) sell (0.915) either (0.915) show (0.913) end (0.908) reports (0.904)
using	having (0.937) making (0.930) taking (0.918) getting (0.908) being (0.893) increased (0.891) about (0.881) within (0.874) then (0.872) at (0.870)
wall	view (0.869) manufacturing (0.857) judge (0.847) defense (0.839) different (0.836) currency (0.835) british (0.834) land (0.833) press (0.808) contracts (0.805)
wednesday	monday (0.967) sales (0.943) part (0.926) 1985 (0.926) university (0.924) low (0.920) program (0.918) office (0.915) december (0.914) budget (0.911)
week	year (0.958) month (0.953) day (0.938) crash (0.933) suit (0.929) group (0.929) settlement (0.928) quarter (0.928) value (0.927) plan (0.926)
weren't	are (0.975) didn't (0.949) had (0.947) ended (0.946) probably (0.941) could (0.921) might (0.918) was (0.915) is (0.914) isn't (0.907)
will	though (0.871) probably (0.825) rose (0.821) i'm (0.816) rather (0.813) are (0.807) disclosed (0.805) could (0.802) also (0.797) once (0.796)
william	retail (0.927) really (0.912) los (0.909) brief (0.885) buying (0.885) treasury (0.880) chicago (0.867) net (0.864) nation's (0.858) bush (0.855)
workers	jobs (0.957) orders (0.954) goods (0.948) inflation (0.947) employees (0.946) america (0.942) concerns (0.938) earnings (0.938) units (0.938) chairman (0.936)

The dendrogram is illustrated in figure C.1 below.

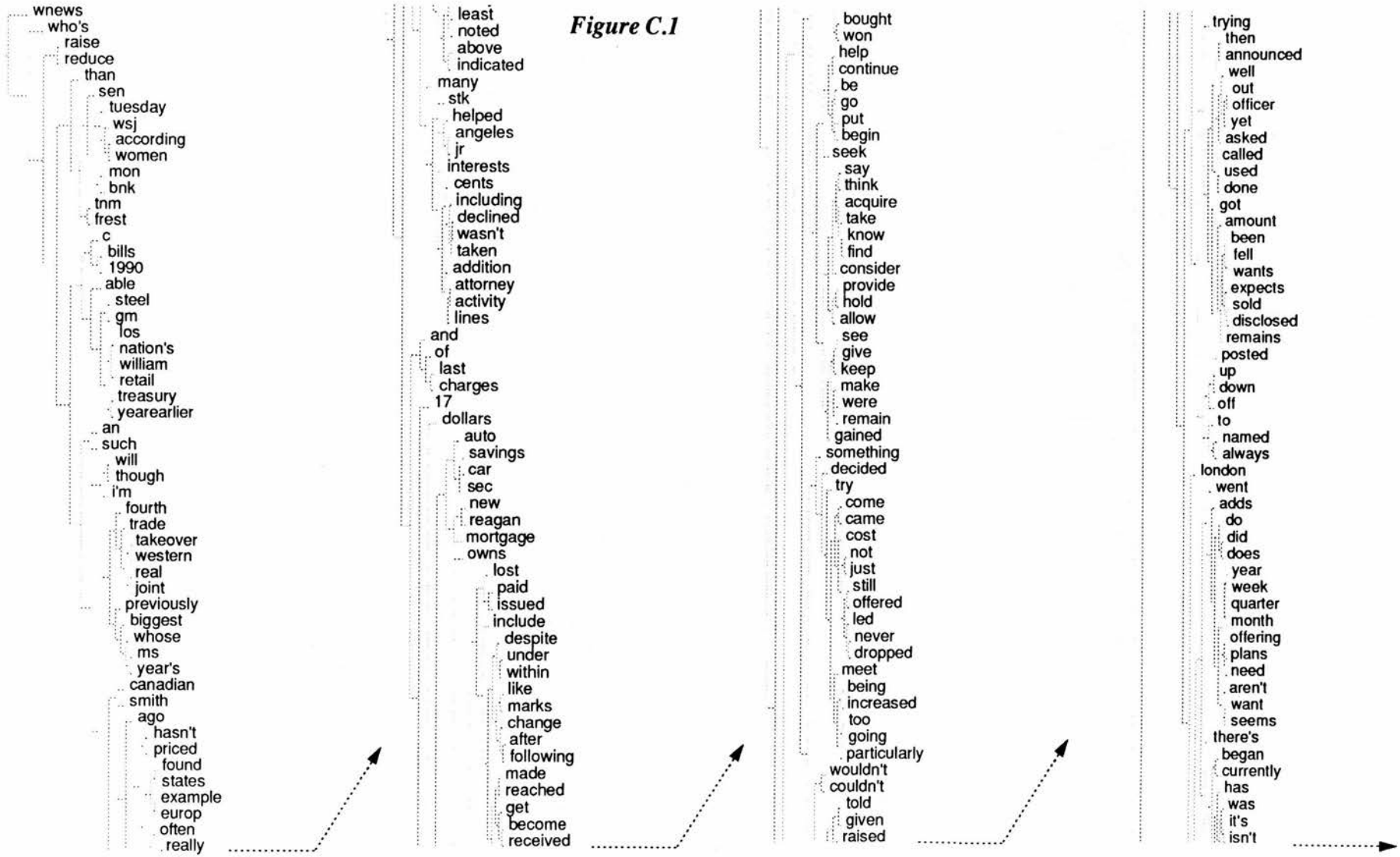

wnews
who's
raise
reduce
than
sen
tuesday
wsj
according
women
mon
bnk
tnm
frest
c
bills
1990
able
steel
gm
los
nation's
william
retail
treasury
yearearlier
an
such
will
though
i'm
fourth
trade
takeover
western
real
joint
previously
biggest
whose
ms
year's
canadian
smith
ago
hasn't
priced
found
states
example
europ
often
really
least
noted
above
indicated
many
stk
helped
angeles
jr
interests
cents
including
declined
wasn't
taken
addition
attorney
activity
lines
and
of
last
charges
17
dollars
auto
savings
car
sec
new
reagan
mortgage
owns
lost
paid
issued
include
despite
under
within
like
marks
change
after
following
made
reached
get
become
received
bought
won
help
continue
be
go
put
begin
seek
say
think
acquire
take
know
find
consider
provide
hold
allow
see
give
keep
make
were
remain
gained
something
decided
try
come
came
cost
not
just
still
offered
led
never
dropped
meet
being
increased
too
going
particularly
wouldn't
couldn't
told
given
raised
trying
then
announced
well
out
officer
yet
asked
called
used
done
got
amount
been
fell
wants
expects
sold
disclosed
remains
posted
up
down
off
to
named
always
london
went
adds
do
did
does
year
week
quarter
month
offering
plans
need
aren't
want
seems
there's
began
currently
has
was
it's
isn't

Figure C.1

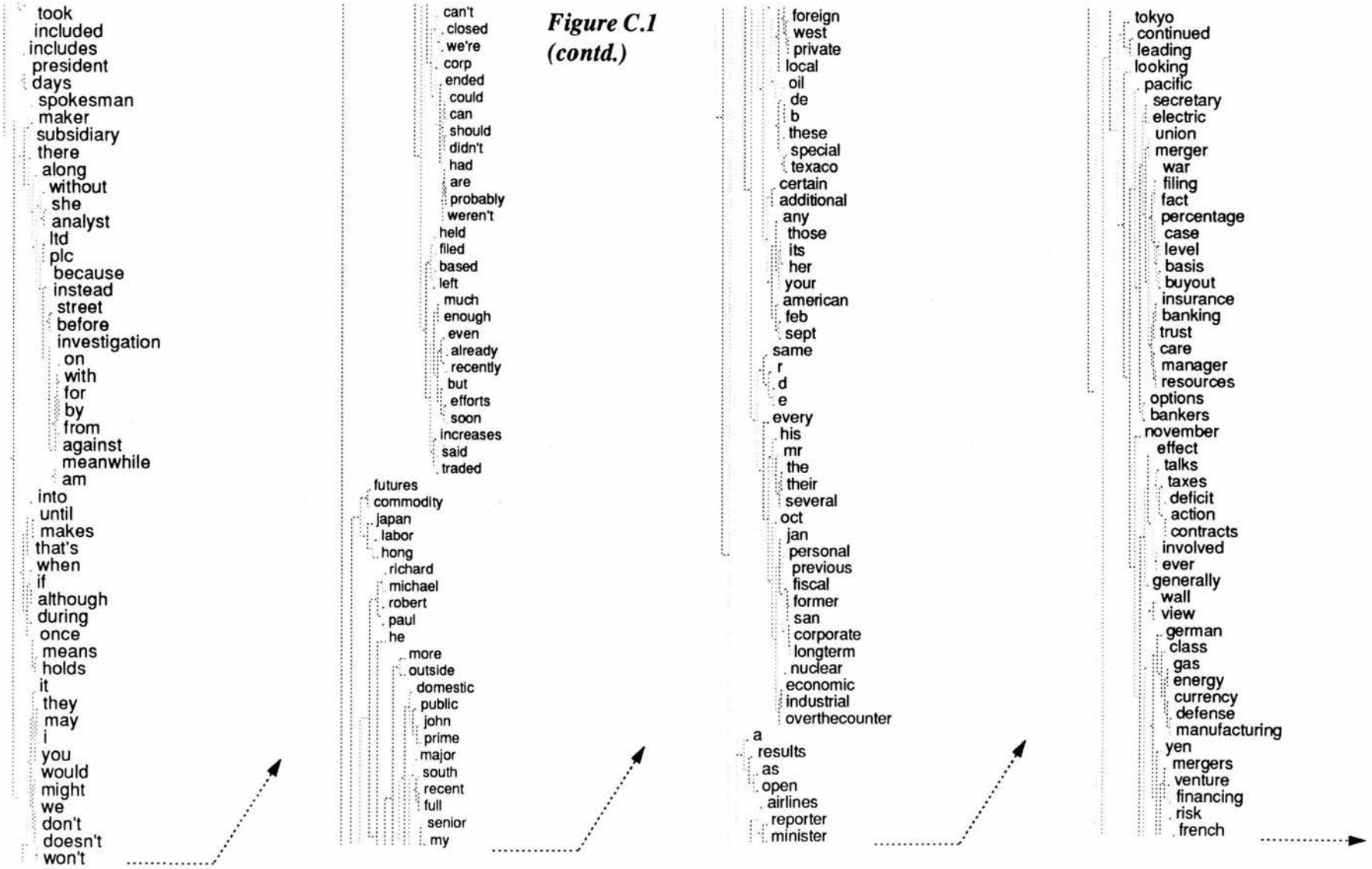


Figure C.1 (contd.)

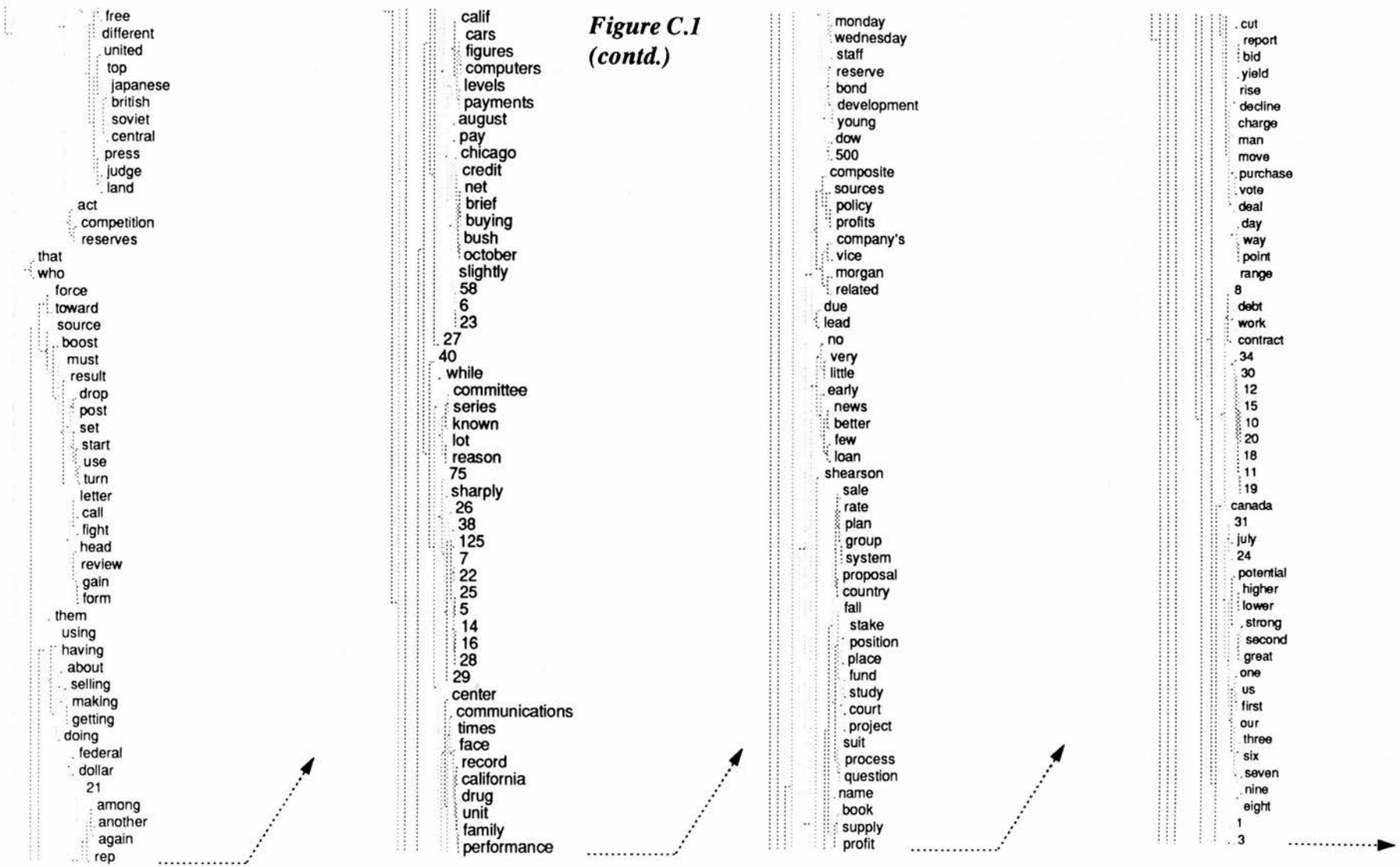


Figure C.1 (contd.)

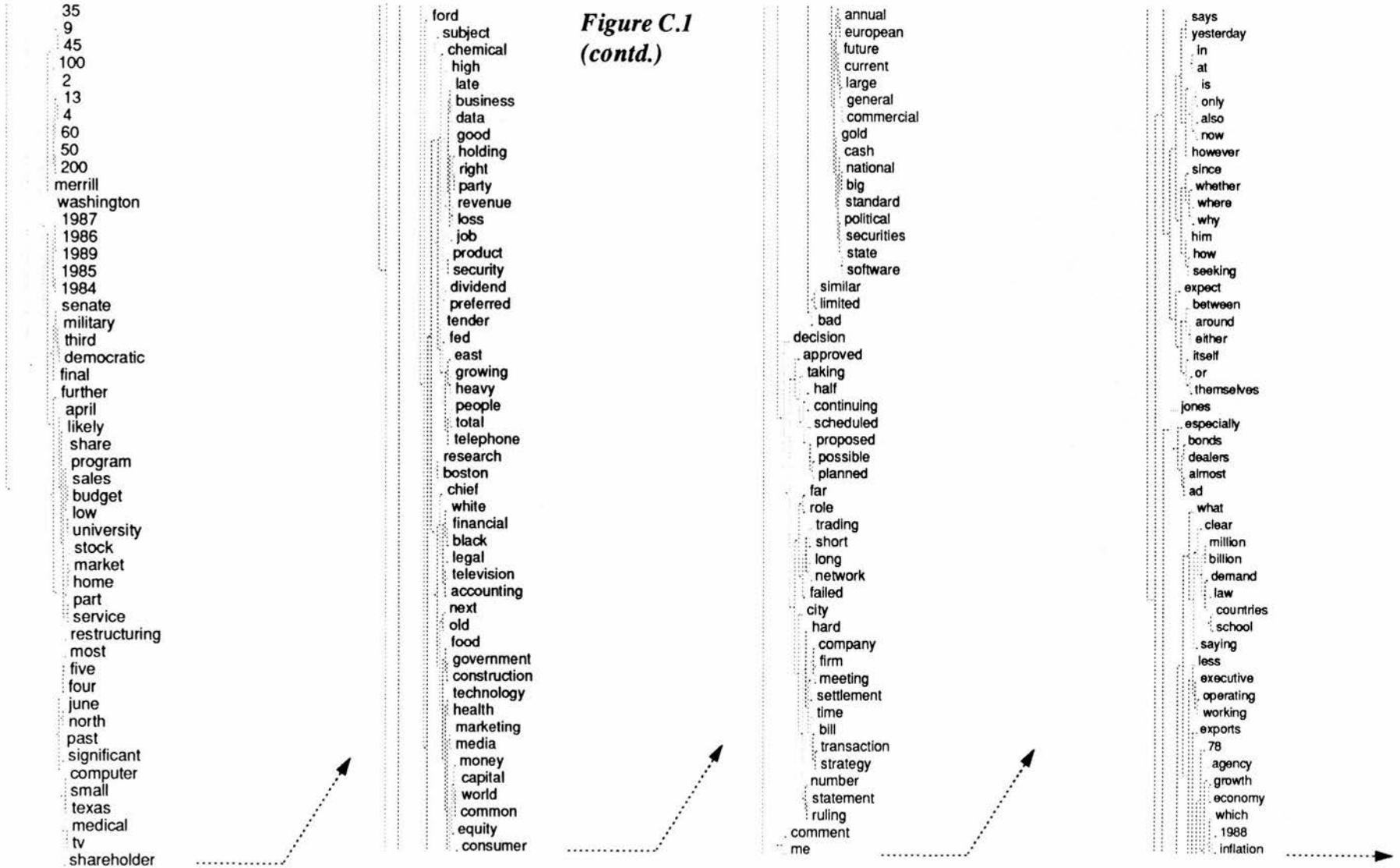


Figure C.1 (contd.)

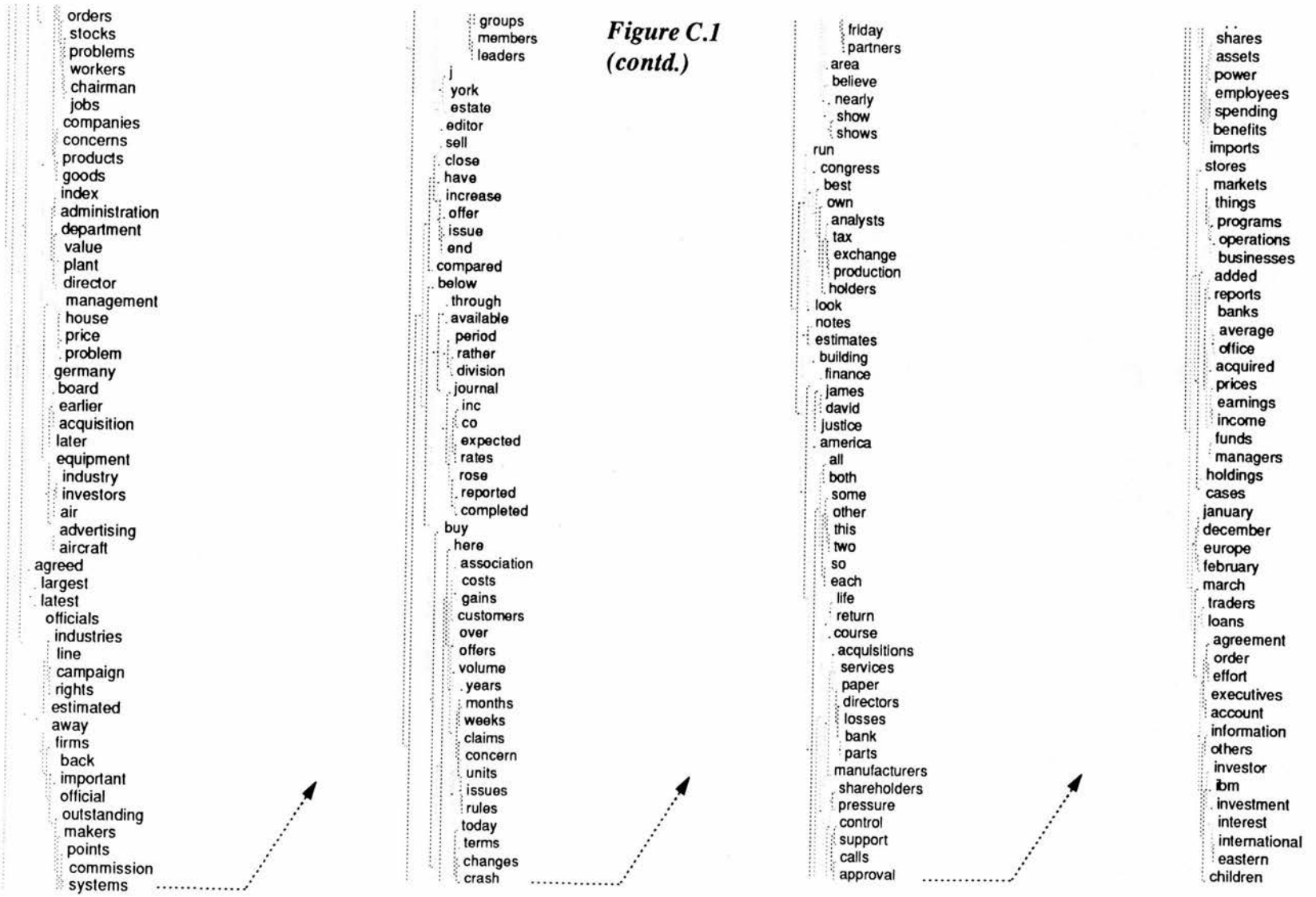


Figure C.1 (contd.)

APPENDIX D

On the following pages, two papers related to the work described in this thesis are reproduced.

The first of these appeared in the Proceedings of the Seventh Conference of the European Chapter of the Association for Computational Linguistics. It is reproduced here with the permission of the publisher, the Association for Computational Linguistics.

The second paper is an unpublished manuscript written in collaboration with John Bullinaria, of the Centre for Speech and Language, Department of Psychology, Birkbeck College, London.

Grouping Words Using Statistical Context

Christopher C. Huckle *
Department of Psychology,
7 George Square,
Edinburgh EH8 9JZ,
Scotland,
U.K.
cch@castle.ed.ac.uk

Abstract

This paper describes the use of statistical analyses of untagged corpora to detect similarities and differences in the meaning of words in text. This work is motivated by psychological as well as by computational issues. The limitations of the method of cluster analysis in assessing the success of such analyses are discussed, and ongoing research using an alternative unsupervised neural network approach is described.

Introduction

There has been considerable recent interest in the use of statistical methods for grouping words in large on-line corpora into categories which capture some of our intuitions about the reference of the words we use and the relationships between them (*e.g.* Brown et al., 1992; Schütze, 1993).

Although they have received most attention from within computational linguistics, such approaches are also of interest from the point of view of psychology. The huge task of developing concepts of word meanings is one that human beings readily achieve; we are all generally aware of the similarities and differences between the meanings of words, despite the fact that in many cases these meanings are not amenable to rigourous definition. Whilst supervision may enable children to learn the meanings of a limited number of common words, it seems extremely unlikely that the greater part of our understanding of word meanings is achieved in this way. Experimental evidence shows (Harris, 1992) that the occurrence of words in young children's language is strongly influenced by the appearance of those words in the speech they hear around them, and it may be that this process continues indefinitely. Such a process would seem to be particularly important when accounting for our understanding of *abstract* words, such as 'similar' and 'justice', which lack concrete referents. Despite our difficulty in being able to provide clear definitions for such words, we have strong intuitions about their usage and can readily categorize them on the basis of similarity in meaning. This process of developing concepts for abstract words is one which psychological research has tended to ignore.

This situation suggests that the learning of the meanings of many words, and their relation to the meanings of other words, may be achieved in an unsupervised fashion, and that our ability to develop a categorization for words may be driven, at least in part, by structure latent in the language being learned. Recent work in computational linguistics which makes use of statistical methods to cluster words into groups which reflect their meaning is attractive in this context as it potentially provides a means for developing conceptual structure without supervision, without giving any prior information about the language to the system, and without making *a priori* distinctions between concrete and abstract words.

Supervision and knowledge of syntax (much useful information about which, as Finch and Chater (1992) have argued, is also contained in simple distributional statistics) are two additional factors which are likely to assist in the process of developing concepts of word meanings. However, by focusing on the single, intralinguistic, source of information provided by the language data alone, we may be able to obtain useful insights regarding its influence on our conceptual structure.

Approaches to Semantic Clustering

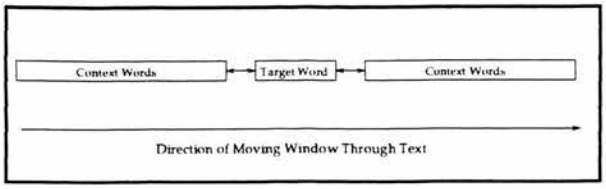
A number of analyses were carried out on text corpora to examine the sorts of semantic groupings that can be achieved using simple statistical methods. Using an approach similar to that of Brown et al. (1992), each 'target word'[1] w_i in the corpus was represented as a vector in which each component j is the probability that any one

*The author is supported by the Carnegie Trust for the Universities of Scotland

[1] For convenience, target words were taken as the n most frequent words in the corpus, with n often equal to 1000

word position in a 'context window' will be occupied by a 'context word' w_j, given that the window is centred on word w_i. The length of the window used can be varied. The basic outline of the moving window used is shown in figure 1. As figure 1 indicates, the portion of the moving window in which the context words are contained may exclude a small number of word positions immediately adjacent to the target word. This is to weaken the effects of syntax, although the analyses described here do not make use of this facility. Following the creation of these vectors, heirarchical cluster analysis was carried out over them, using Euclidean distance between vectors as a similarity metric. Analyses were also carried out in which, as with Finch and Chater (1992), the distance metric used was the Spearman Rank Correlation coefficient. The approach described here differs from that of Brown et al. (1992) in that context words both preceding and following the target word are considered (although information about the *ordering* of the context was not used), and in that Euclidean distance, rather than average mutual information, is used for clustering.

Figure 1: Design of the Moving Window



Each of the methods described here represents each target word in the same manner, regardless of the syntactic or semantic designation which might conventionally be assigned to it. Thus any differences or similarities between words must be detected purely from the statistics of the usage of the words, which are in turn determined by the characteristics of the contexts in which they occur.

Results

The methods outlined above were used to cluster words appearing in the Lund corpus (470,000 words), a corpus created from issues of the Wall Street Journal (1.1 million words), and a corpus created from the works of Anthony Trollope (1.7 million words).

Initial analyses were carried out on the Lund and Trollope corpora using a short window length of only one word position either side of the target word. That is, target words were represented by vectors whose components reflected the (bigram) statistics of occurrence of context words at the word position immediately preceding the target word or immediately following the target word. Whilst it seems reasonable to suppose that children acquiring word meanings would be able to make use of more than this limited amount of context information, the analyses were carried out to investigate performance of the system under such crude conditions.

It was found on examination of the dendrograms resulting from the cluster analyses that even using this extremely impoverished source of information about the target words did permit a limited number of semantically coherent groupings of words to be created. The members of some of these groups were selected following inspection of the relevant dendrograms and are listed in table 1. Despite the existence of the groupings shown in table 1 and a small number of others like them, they represent only a small proportion of the 1000 target words subjected to the analysis. Besides those shown above, a number of other types of groupings were evident which appeared to reflect *syntactic* rather than more specific semantic characteristics. This is perhaps not surprising if one regards the problem of grouping words on the basis of similarity as one of prediction; given statistical information only about those words immediately adjacent to a particular target word, it may be possible to say with reasonable confidence that the target word is a noun, a verb, or an adjective, but information about wider context is likely to be needed in order to provide more specific predictions about the *particular* noun, verb, or adjective in question. Since this information is not present, the dendrograms resulting from the analysis show groupings of prepositions, adjectives, verbs, and so on. Also present are groups of words whose members all commonly precede or follow a particular particle.

Table 1: Semantic Groupings

Possible Designation of Group	Group Members
Mental States (Lund Corpus)	want, wanted, tried, went, decided, think, thought, hope, believe, knew, feel, felt, expect, wish, forget.
Days of the Week (Lund Corpus)	friday, thursday, saturday, sunday, monday, wednesday, tuesday.
Measures (Lund Corpus)	ninety, pounds, years, days, minutes, hours, double, miles.
People (Lund Corpus)	boy, girl, man, woman.
Numbers (Trollope Corpus)	six, twelve, twice, twenty, two, three, four, ten, five, seven.
Units of Time (Trollope Corpus)	months, years, days, hours, o'clock, times.
Parts of the body (Trollope Corpus)	arm, mouth, pocket, arms, chair, sister, thoughts, feet, eye, heart, father, face, head, eyes, hand, ears, hands, bosom.
Human Family Members (Trollope Corpus)	aunt, mind, uncle, husband, cousin, mother, daughter, brother, niece.

Further analyses were carried out in which the length of the context window was extended to 5

words either side of the target word. The dendrograms resulting from these analyses did not show any marked improvement over those obtained from the earlier analyses, and even when the window length was increased to 25 words each side of the target word, clear differences were not easy to detect from the dendrograms, although the sorts of groupings noted earlier were still identifiable.

Future Directions

The use of cluster analysis and related techniques has been popular for presenting the results of recent statistical language work within computational linguistics. However, such methods clearly have a number of limitations. Firstly, it is difficult to compare dendrograms rigourously, which means that it can be difficult to determine which of a number of alternative approaches or sets of parameters is turning out to be the most successful. Secondly, the lack of an objective measure of the clusters obtained means that assessments of the success of a particular technique for categorizing language may well be unreliable; it is quite possible to focus on the attractive looking groupings revealed in a dendrogram whilst ignoring what may be a very large number of less attractive ones.

These criticisms arise largely because cluster analysis is a purely descriptive statistical method, and strongly suggest that alternative methods must be found which can provide a more objective measure of the success of the technique being used. Of these, word sense disambiguation is attractive. Since we can obtain from native speakers an assessment of the correct senses of target words in different contexts, we do have a means for determining how often a particular technique is able to give the correct sense for a particular target word. In other words, the evaluation of a native speaker can potentially be used to assess performance each time the system encounters a target word in context and assigns that word to a particular sense class. Whilst such assessments might also be applicable to the analysis of dendrograms, word sense disambiguation is of interest since it constitutes the task that continually meets human language users when reading text or listening to speech.

For these reasons, current work is focusing on the problem of disambiguating words given statistical context. To achieve this, an unsupervised competitive neural network is being used. This has several features which appear to be desirable. Firstly, as in the human case, learning proceeds on-line, without any need for a separate stage of statistical analysis. Such a system has the potential to begin developing clusters from the very first exposure to the linguistic input, and the clusters into which the input words are placed evolve continuously during the learning process. Thus one can usefully examine the state of the clusters at any point during learning. Secondly, it is straightforward to allow any given word to be clustered into as many separate clusters as the system dictates (subject to the maximum number of output units available). Thus, the neural network approach, unlike that described above, has the potential to allow separate senses of a word to be distinguished on the basis of their context. This is not to say that non-neural network approaches could not permit a word to belong to more than one cluster (*e.g.* Pereira et al., 1993), but rather that this is a very natural and attractive consequence of using the unsupervised neural network approach.

At present, work is being undertaken to examine how well a simple competitive neural network can perform on such a task. Preliminary work has been undertaken using a simple competitive neural network similar to that described by Finch and Chater (1992). Unlike them, though, provision was made for presenting words along with context during the test phase as well as the training phase. This potentially allows disambiguation performance to be examined at any time. Initial work using the very simple artificial corpus devised by Elman (1988) has been encouraging, with the network demonstrating near-perfect performance in distinguishing between nouns and verbs in the corpus.

References

Peter F. Brown, Vincent J. Della Pietra, Peter V. deSouza, Jenifer C. Lai, and Robert L. Mercer. 1992. Class-based *n*-gram models of natural language. *Computational Linguistics*, 18(4):467–479.

Jeffrey L. Elman. 1988. Finding Structure in Time. CRL Technical Report 8801. Center for Research in Language, University of California, San Diego.

Steven P. Finch and Nicholas J. Chater. 1992. Bootstrapping syntactic categories. In *Proceedings of the 14th Annual Conference of the Cognitive Science Society Of America*, pages 820–825. Bloomington, Indiana.

Margaret Harris. 1992. *Language Experience and Early Language Development.* Lawrence Erlbaum, Hove, U.K.

Fernando Pereira, Naftali Tishby, and Lilian Lee. 1993. Distributional Clustering of English Words. In *Proceedings of the Association for Computational Linguistics*, volume 31, pages 183-190. ACL.

Hinrich Schütze. 1993. Part-of-speech-induction from scratch. In *Proceedings of the Association for Computational Linguistics*, volume 31, pages 251–258. ACL.

Modelling Lexical Decision Using Corpus Derived Semantic Vectors in a Connectionist Network

John A. Bullinaria

Centre for Speech and Language
Department of Psychology
Birkbeck College, Malet Street
London WC1E 7HX, UK

email: johnbull@ed.ac.uk

Christopher C. Huckle

Neural Networks Research Group
Department of Psychology
University of Edinburgh, 7 George Square
Edinburgh EH8 9JZ, UK

email: cch@castle.ed.ac.uk

Abstract

We discuss the use of non-binary corpus derived semantic vectors in connectionist models of lexical decision. The co-occurrence statistics of words in large corpora allow the generation of vectors whose distribution correlates with the perceived semantic relatedness of the words. Connectionist models of the mapping from phonology or orthography to random binary semantic vectors allow the simulation of lexical decision reaction times that show patterns of semantic and associative priming similar to those found experimentally with human subjects. We consider the problems of extending these connectionist models to deal with the non-binary corpus derived vectors. We do find significant lexical decision priming predicted by distances in the semantic vector space, but the reaction times are very noisy. Averages over many words and/or many networks are required for the relationships to become clear. The question of associative priming remains open.

Introduction

Lexical decision (i.e. the task of deciding whether a given string of letters or phonemes is a real word) is widely used in psychological experiments to investigate the processes and representations employed in basic human language processing (e.g. Neely, 1991; Shelton & Martin, 1992; Moss et al., 1995). Of particular interest is the study of priming (i.e. the effect by which response is speeded through prior presentation of certain related words). The fact that we observe facilitation by semantically related words (e.g. 'jump' primes 'leap') suggests that the lexical decision process taps into some underlying semantic representation and that lexical decision experiments can be designed to explore these representations. However, priming is also found to be produced by words that are associated but not semantically related (e.g. 'pillar' primes 'society'). It is not yet clear if such associative priming needs to be explained by a different mechanism to semantic priming or if it is inherent in the properties of the same semantic representations.

Recently, connectionist models of the lexical decision process have been constructed that account for many aspects of semantic and associative priming (Plaut, 1995; Bullinaria, 1995). The semantic priming arises due to the overlap of distributed semantic vectors and the associative priming arises due to word co-occurrence during learning. However, these models have been based on hand-crafted and/or random binary semantic vectors. Other researchers (e.g. Lund et al., 1995) have derived semantic vectors from large text corpora and have suggested that the distances between words in this semantic vector space can account for the experimental priming results. In this paper we put these two approaches together and investigate the properties of connectionist lexical decision models based on corpus derived semantic vectors. In particular, we question whether it is really possible to obtain useful results *without* considering the two approaches together.

Modelling Lexical Decision

Given the experimental evidence that semantics has an effect on lexical decision reaction times, it is natural to assume that the time taken to activate the appropriate semantic representation provides at least one factor in the lexical decision process. Within the conventional connectionist framework we model this by choosing (simplified) representations for the orthography/phonology and semantics and setting up a network to map between them. Plaut (1995) chose to use a recurrent network trained with continuous back-propagation through time. Bullinaria (1995) used a cascaded feed-forward network trained in a similar

manner. Both approaches led to similar (though experimentally distinguishable) patterns of reaction times and priming. Here we shall adopt the Bullinaria (1995) framework.

For simplicity, we restrict ourselves to mono-syllabic words and represent phonology by having one unit for each possible onset, vowel and offset phoneme cluster. Each word then has three phonological input units activated. Since the phonology to semantics mapping in English is essentially random (ignoring morphological effects) it is not unreasonable to represent the semantics by random binary vectors with the interpretation that activated units correspond to the small number of relevant semantic micro-features (cf. Plaut & Shallice, 1993). The network will then require a sufficiently large layer of 'hidden units' in order to handle the random and non-linearly separable associations between this phonology and semantics.

Since we are aiming to model reaction times (RTs), it makes sense to think in terms of activation cascading through the network (e.g. McClelland, 1979) as in recurrent networks rather than the typical one pass approach of standard feed-forward networks. To simulate this we discretize the time and at each time slice t we take:

$$\mathrm{Out}_i(t) = \mathrm{Sigmoid}(\mathrm{Sum}_i(t))$$

$$\mathrm{Sum}_i(t) = \mathrm{Sum}_i(t-1) + \lambda \sum_j w_{ij}\mathrm{Prev}_j(t) - \lambda \mathrm{Sum}_i(t-1)$$

with the output $Out_i(t)$ of each unit i the usual sigmoid of the sum of the inputs into that unit at that time. The sum of inputs $Sum_i(t)$ is given by the existing sum at time $t-1$ plus the additional weight w_{ij} dependent contribution fed through from the activation $Prev_j(t)$ of the previous layer and a natural exponential decay of activation depending on some time scale λ.

There are now two broad approaches to training the network. The quick way is to note that, as long as we have static inputs, the asymptotic state of the above equations reduce to:

$$\mathrm{Out}_i(t_\infty) = \mathrm{Sigmoid}(\mathrm{Sum}_i(t_\infty))$$

$$\mathrm{Sum}_i(t_\infty) = \sum_j w_{ij}\mathrm{Prev}_j(t_\infty)$$

which are the equations for a standard non-cascaded feed-forward network. It follows that, if we only require the network to produce correct outputs for individual words, we can simply train this asymptotic state using a standard gradient descent algorithm (such as back-propagation). The resultant trained network can then be used in a cascaded fashion to extract the RTs. If, however, we want the network to respond efficiently to sequences of words, we need to train *during* the cascading process so the network can *learn* to make quick transitions from one set of activations to another. The network can still be trained using a standard gradient descent procedure to modify the weights w_{ij} iteratively so that the output activation errors are reduced. However, for each input word, we now need to repeat this process over many time slices as the network settles into a stable state. If we present the training words in random order, and keep the time parameter λ and learning rate ε sufficiently small that large fluctuations in the weights and activations do not occur, then the network eventually learns to produce the correct outputs for any word without any resetting after the previous word.

Reaction times can then be defined in terms of time slices in a number of ways. We could simply take the time required for the network to settle into a stable output semantic state (as in Plaut, 1995). Alternatively, we could attempt to be more explicit about modelling the lexical decision process by timing the consistency checking between the input phonology and the phonology produced by allowing activation to flow from phonology to semantics and back to phonology. This simple 'activate and check' mechanism was shown in Bullinaria (1995) to be able to provide a reliable method of performing lexical decision in this kind of model, whereas details of the pattern of semantic activation alone were not sufficient. Finally, we could argue that the semantic output activations need to drive some later decision process, and that we can ignore the details of this process and take the time required for the integrated output activations to reach some threshold. This may be feasible if all the semantic vectors had equal numbers of fully activated units, but if different words have unequal numbers of activated units (e.g. to represent word concreteness as in Plaut & Shallice, 1993) or if we have non-binary activations (as we shall consider later), then this approach makes less sense. In the following we shall consider both the settling and consistency checking times, and compare their results. In each case we first activate the network for the prime word and then, without resetting the activations, present the target word and measure the RT.

In this framework, semantic priming arises naturally due to overlap of the semantic vectors. If the network activations due to the prime word are already close to that which will be activated for the target word, then it will take fewer time slices for the target word to be activated from this state than if it were starting from the activation pattern of some unrelated control word. Associative priming may also be caused by properties of the semantic vectors, but it has been shown explicitly in connectionist models (Moss et al.,

1994; Plaut, 1995; Bullinaria, 1995) how the facilitation can arise purely due to co-occurrence of words in the training data. If, for example, 'society' follows 'pillar' much more often during training than would be expected by chance, then it is not surprising that an efficient learning system will be able to make use of this fact to speed its response times.

Corpus Based Semantic Vectors

As noted above, there has been considerable interest recently in using statistical vectors derived from large text corpora to represent the contexts in which words occur, and thus to provide a representation for their semantics. The idea is that words which occur in similar contexts will tend to have similar meanings. Advantages of this approach include the elimination of the need for semantic features to be supplied by hand (e.g. as in Plaut & Shallice, 1993) and the parsimony of being able to use that statistical structure latent within the language itself. Work of this type has often been pursued from the perspective of computational linguistics (e.g. Brown et al., 1992). However, Huckle (1995) has recently noted that it is also of importance in exploring psychological questions concerning human acquisition and representation of word meanings, and has discussed such methods in conjunction with the use of neural networks.

The general approach is to represent each word's distributional context using a vector of probabilities obtained by 'reading' large samples of natural language text. For every 'target word' $word_p$ being represented, each component of its vector contains the probability that some other 'context word' $word_q$ will occupy a particular relationship to $word_p$ in the text. This relationship is typically one which concerns the physical distance between the two words. Once vectors of this general type have been obtained, the distances between them can be calculated to reveal similarities between word meanings and suitable transformations may be applied to provide useful semantic representations for neural networks (e.g. Schütze, 1993). Alternatively, the inter-word distances in the semantic space can be used directly as a basis for investigating psychological phenomena such as semantic priming (e.g. Lund et al., 1995).

To enable reliable statistics to be obtained, a large corpus must be used. Here, we used a 10,000,000 word corpus taken from issues of the Wall Street Journal published in 1988 and 1989. Since Zipf's law informs us that the probabilities for less frequent words in a corpus of this size will rapidly become less reliable (Zipf, 1935), we restricted ourselves to the use of the most high frequency items. The most frequent 1000 words in the corpus were taken as our target words, while the most frequent 200 words were used as our context words.

This still left us with many possibilities for calculating useful semantic vectors. For each of the 1000 target words, we derived a 200 dimensional vector in which each component contained the probability that a particular context word would occur within a window of two words to the left of the target word in the corpus. Similar vectors were also calculated in which each component was the probability of a particular context word occurring within a window of two words to the right. The window length of two words was chosen following exploratory work which suggested this would be optimal for capturing a word's semantic context. In calculating the co-occurrence probabilities from the word counts, we normalized for the frequency of the appropriate target word in each case to remove the bias this would otherwise have had on the vector components.

Finally, following further exploratory work, we concluded that the best overall semantic representations were obtained by simply concatenating our left and right context vectors to give a 400 dimensional vector of probabilities for each of the target words. Our final approach to representing the target words is thus similar to that adopted by Lund et al. (1995).

The Combined Model

Unfortunately, using our corpus based vectors in our lexical decision model was not a totally straightforward matter. Our first problem was that, if we are going to train our networks in a reasonable amount of time, we need to keep the networks as small as possible, which in turn means minimising the dimensionality of our semantic vector space. For our purposes, principal component analysis proved a convenient procedure. The 400 dimensional semantic vectors were projected onto the 30 dimensional sub-space containing the maximum variance. This provided much lower dimensional vectors with relatively little loss of information and the added advantage that we lost a lot of the noise in the process. The inter-word distances in the original space and the sub-space correlated well (Pearson $r = 0.94$).

A second problem arose with the standard use of sigmoidal activation functions in our network. Given that a sigmoid transformation leads to a distortion of the distances in semantic space and that distortion depends on an essentially arbitrary scale factor, we decided to use a linear activation function for the

Prime Set	Distances	Settling Times	Consistency Times
C1	56.1 (4.2)	669 (44)	709 (77)
C2	45.9 (4.3)	655 (58)	699 (75)
C3	39.9 (5.3)	638 (47)	676 (83)
P3	9.6 (5.4)	591 (72)	586 (91)
P2	8.6 (5.2)	586 (71)	569 (93)
P1	7.1 (4.4)	581 (78)	570 (101)

Table 1. The simulated primed RTs compared with distances in semantic space.

network outputs. Since the distribution of vector components was rather skewed towards small values anyway, which corresponds to the central linear region of the sigmoid rather that the saturated extremes, this difference is probably not crucial.

What might be crucial however, is the decision to use the non-binary components that come out of the corpus analysis rather than attempting to convert them into the binary form commonly used in connectionist systems. Since we already know that we can get semantic and associative priming using random binary semantic vectors, it seemed more interesting to investigate the more ambitious case of non-binary vectors. Moreover, in the human case, it would be natural to consider our semantic representation to be a 'hidden representation' that is learnt by the brain, and it is rare to find hidden representations developing binary values in connectionist models. We shall see later how our networks behave rather differently when based on real, rather than binary, outputs.

The next thing we have to consider is that the Wall Street Journal is a rather atypical source of the English language. Certain word pairs (such as 'dow jones' and 'wall street') occur together much more frequently than in normal English, and other words (such as 'bush' and 'ford') are often used in atypical ways. To avoid possible artefacts that these words may cause, we simply removed them from our network training set. We also removed 16 homographs which would clearly have problematic semantic vectors.

In Bullinaria (1995), the mapping from phonology to semantics was discussed. Here (as in Plaut, 1995) we consider the mapping between orthography and semantics. Given our simplified abstract input representations and the regularity of the orthography to phonology relationship, the distinction is unlikely to be crucial, but it should be kept in mind. As with other models that map to semantics, the randomness of the mapping means that in order to train the network in a reasonable amount of time, we need to restrict the number of training words, which allows us to use less hidden units. It then follows that we have to reduce the size of the input space so that the word distribution does not become unnaturally sparse. To this end we restricted ourselves to monosyllables with orthography made up of the most common onset consonant clusters (30), vowel clusters (18) and offset consonant clusters (38) plus two units to code for the presence or absence of a final 'e'. Our set of 1000 target words contained 270 words consistent with all the above restrictions. These all occurred within the most frequent 993 words in the corpus and with an occurrence of at least 1297. Finally, we set the arbitrary scale and origin of the semantic space so that the mean activation of each semantic unit was zero and the overall standard deviation was 0.14, with maximum component 1.95 and minimum component –1.27.

We trained our main network on these words using back-propagation of errors on the asymptotic output patterns (with 270 hidden units, sum squared error measure, learning rate $\varepsilon = 0.01$, no momentum, 75000 epochs). We also attempted to train a similar network throughout the cascading process with no resetting of activation between words (150 time slices per word, $\lambda = 0.1$, sum squared error measure, learning rate $\varepsilon = 0.0001$, no momentum). The RTs were then extracted as described above, except that we used a reduced $\lambda = 0.01$ to give a more accurate approximation to the continuous process. The settling RTs were defined as the number of time slices required for all output activation changes per time slice to fall below 0.0001. The consistency RTs were the number of time slices required for the total difference between the input and output phonology activations to fall below 0.1.

Semantic Priming

Needless to say, we checked that our closest semantic vectors (defined in terms of Euclidean distance) did actually correspond to words that were semantically related ('will' to 'would' 2.1, 'three' to 'four' 2.1) and that the most distant words really were unrelated ('lot' to 'past' 73.7, 'same' to 'try' 64.6), though in this paper we shall not attempt to match our network results to real lexical decision experiments. For that we really need a much larger number of much larger networks and vectors derived from much more representative corpora.

We have noted already that semantic priming has been found before in networks such as ours (Bullinaria, 1995). In these binary target networks, the semantic Euclidean distances were all $\tilde{A}2 \sim 1.41$ for semantically related items and $\tilde{A}6 \sim 2.45$ for un-related items, yet there was still a large distribution of RTs and degrees of priming. Clearly factors other than semantic distances were at work. Inevitably, the RT will be determined by the unit i needing to swap values that is the slowest. Looking at the above cascade equations we see that, to first approximation, the number of time slices to move from the prime to settled target output will be given by the difference of the final $Sum_i(t\_, prime)$ and $Sum_i(t\_, target)$ divided by the average step size which will also be related to $Sum_i(t\_, target)$. We can not use this as an easy way to predict the RTs, because determining the average step size is harder than actually running the network, but it does tell us something useful about the RTs. The problem with sigmoids and binary output targets is that the sigmoids saturate, which means that very large random variations may arise in the Sum_i's during learning without changing the actual network outputs very much and hence without being constrained by the training algorithm. Given that the values of the Sum_i's have such a big influence on the RTs, it is no wonder that the RTs are so noisy. With our non-binary semantic vectors we have no output sigmoids and the output activations are the Sum_i's themselves which are learnt to be particular values. Thus we may expect to suffer less seriously from random effects than the binary case. But is this true (because we still have sigmoids at the hidden layer) and does it allow the simulated priming results to correlate with the distances between our semantic vectors?

There are many ways to illustrate our network priming results. We begin by looking at the RTs for each of our 270 words when primed by the three closest words in semantic space (sets P1, P2, P3) compared to the RTs when primed by the three furthest words (sets C1, C2, C3). The mean distances and RTs are shown in Table 1 (with standard deviations in brackets). The first thing to notice is that both forms of simulated RT show faster times for the prime sets (P's) than the control sets (C's), so we do find semantic priming (and this is highly significant, $p < 0.0001$). However, it is clear from the standard deviations that the RTs are very noisy and that there is no simple relationship between the Euclidean distances and the RTs. Indeed, if we take what we would expect to be the maximum priming result, i.e. the differences in RTs between control prime set C1 and semantically close prime set P1, we find the effect is not always even in the right direction. Figure 1 shows the distribution of priming for our two RT approaches. Not only do both approaches show some negative priming, but they also fail to agree on which words this happens for. For the settling RTs we have mean 88 (standard deviation 76), minimum –72, maximum +293 For the consistency RTs we have 140 (101), –135, +429. If we average over all three sets of primes and control primes, the priming is only slightly more reliably positive: 68 (48), –69, +205 for settling, 119 (61), –38, +322 for consistency. Such distributions of results are also found in human subjects, so this is not a problem in that respect, but it does make it rather difficult to see the precise relationship between the semantic space distances and the network priming results.

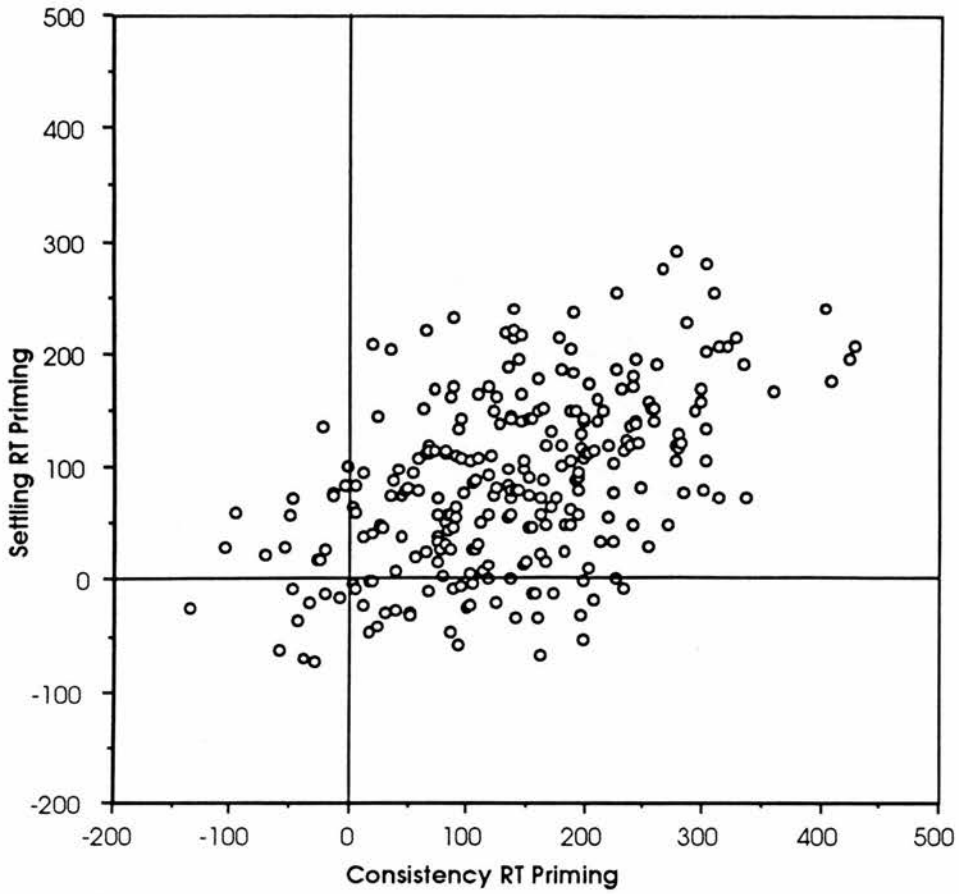


Figure 1. Comparison of priming results for settling and consistency RTs.

We really need to consider many networks in the same way that experimental results are averaged over many subjects. Hence we trained another three networks with different random initial weights and learning rates, and obtained results similar to Figure 1, though the individual word primings did not correlate that well between networks (Pearson r ~ 0.5). Averaging over the four networks and three primes per word gave priming of 62 (37), –37, +169 for settling, 111 (49), –46, 260 for consistency. These results suggest that averages over sufficiently many networks and/or primes may cancel out the noise leaving clear semantic effects plus any other (e.g. orthographic competitor, RT simulation dependent) effects not inherent in the semantic vectors themselves. Unfortunately, the indications are that a rather large number networks may be required in order to obtain a clear picture of what is going on.

Associative Priming

For random binary semantic vectors, we find that our networks when trained throughout the cascading process are able to use frequent word co-occurrences during training to speed their RTs, and hence exhibit associative priming (Bullinaria, 1995). This is largely because sigmoidal networks with binary output targets have the flexibility to easily make the weight adjustments appropriate for an advantage in RT with little increase the output error. For example, weight changes could result in a large shift (from –15.0 to –5.0 say) in a $Sum_i(t\_)$, with only a small increase (~0.007) in the activation error $Out_i(t\_)$. In our linear non-binary output networks it is much harder to gain an RT advantage in this way, since any significant change to an output $Sum_i(t\_)$ will directly introduce a significant error into the $Out_i(t\_)$. When using the standard cascaded learning approach described above in this case, the output errors that inevitably occur during the word transitions cause weight changes that disrupt the networks' performance to such an extent that it is difficult for the network to learn accurate semantic representations (as checked by testing them on the phonology to semantics mapping). If we artificially reduce the disruption by increasing λ (to 0.5), the network *is* able to learn, and it does show both semantic and associative priming. It thus remains an open question as to whether human brains are able

to use this word co-occurrence mechanism with fixed semantic representations (such as derived from corpus statistics) to achieve associative priming. It is possible that the semantic representations themselves must be subject to adjustment during training in order to do so. It is also possible that the experimental associative priming is already inherent in the 'semantic' vectors without the need for any additional training effects. It may even be that additional connections between associated words (i.e. their semantic micro-feature units) are employed. Clearly, many more network simulations are required to resolve this matter. Of course, it should be noted that the experimental results are far from clear cut either. Separate claims have been made that all associative priming is really semantic priming (Lund et al., 1995), that all semantic priming is really associative priming (Shelton & Martin, 1992), and that both exist in their own right (Moss et al., 1994, 1995).

Conclusions

We have investigated the use of non-binary corpus derived semantic vectors in connectionist models of lexical decision. Our main conclusion is that there *is* a significant relation between distances in corpus derived semantic vector spaces and priming in the connectionist networks that use these vectors, *but* the relation is very noisy. In the same way that experiments need many test words and many subjects to show clear priming results, so do the connectionist simulations. We believe the preliminary combined models presented here show much promise, but clearly further investigation with larger more representative corpora and many larger and more realistic networks are required before we can be sure of the precise relationship between the corpus, network and experimental results.

References

Brown, P.F., Della Pietra, V.J., deSouza, P.V., Lai, J.C. & Mercer, R.L. (1992). Class-based *n*-gram models of natural language. *Computational Linguistics*, **18**, 467-479

Bullinaria, J.A. (1995). Modelling Lexical Decision: Who needs a lexicon? In J.G. Keating (Ed.), *Neural Computing Research and Applications III*, 62-69. Maynooth, Ireland: St. Patrick's College.

Huckle, C. (1995). Grouping Words Using Statistical Context. *Proceedings of the Seventh Conference of the European Chapter of the Association for Computational Linguistics*, 278-280. San Francisco, CA: Morgan Kaufmann.

Lund, K., Burgess, C. & Atchley, R.A. (1995). Semantic and Associative Priming in High-Dimensional Semantic Space. *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society*, 660-665. Mahwah, NJ: Erlbaum.

McClelland, J.L. (1979). On the time relations of mental processes: An examination of systems of processing in cascade. *Psychological Review*, **86**, 287-330.

Moss, H.E., Hare, M.L., Day, P. & Tyler, L.K. (1994). A Distributed Memory Model of the Associative Boost in Semantic Priming. *Connection Science*, **6**, 413-427.

Moss, H.E., Ostrin, R.K., Tyler, L.K. & Marslen-Wilson, W.D. (1995). Accessing Different Types of Lexical Semantic Information: Evidence From Priming. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **21**, 1-21.

Neely, J.H. (1991). Semantic Priming Effects in Visual Word Recognition: A Selective Review of Current Findings and Theories. In D. Besner & G.W. Humphreys (Eds), *Basic processes in reading: Visual word recognition*, 264-336. Hillsdale, NJ: Erlbaum.

Plaut, D.C. (1995). Semantic and Associative Priming in a Distributed Attractor Network. *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society*, 37-42. Mahwah, NJ: Erlbaum.

Plaut, D.C. & Shallice, T. (1993). Deep Dyslexia: A Case Study of Connectionist Neuropsychology. *Cognitive Neuropsychology*, **10**, 377-500.

Schütze, H. (1993). Word Space. In S.J. Hanson, J.D. Cowan & C.L. Giles (Eds), *Advances in Neural Information Processing Systems 5*, 895-902. San Mateo, CA: Morgan Kauffmann.

Shelton, J.R. & Martin, R.C. (1992). How Semantic is Automatic Semantic Priming? *Journal of*

Experimental Psychology: Learning, Memory and Cognition, **18**, 191-210.
Zipf, G.K. (1935). *The Psycho-biology of Language*. Boston: Houghton Mifflin.

DECLARATION

This thesis has been composed by me and the work contained within it is my own.

Christopher Cedric Huckle