Exploration in Automatic Thesaurus Discovery

Article · January 1994

DOI: 10.1007/978-1-4615-2710-7

1 author:

Gregory Grefenstette
Florida Institute for Human and Machine Cognition
152 PUBLICATIONS **3,893** CITATIONS

EXPLORATIONS IN AUTOMATIC THESAURUS DISCOVERY

EXPLORATIONS IN AUTOMATIC THESAURUS DISCOVERY

Gregory GREFENSTETTE
University of Pittsburgh
Pittsburgh, Pennsylvania, USA

Rank Xerox Research Centre
Grenoble, France

KLUWER ACADEMIC PUBLISHERS
Boston/London/Dordrecht

CONTENTS

PREFACE

by David A. Evans
Associate Professor of Linguistics and Computer Science
Carnegie Mellon University

There are several 'quests' in contemporary computational linguistics; at least one version of the 'Holy Grail' is represented by a procedure to discovery semantic relations in free text, automatically. No one has found such a procedure, but everyone could use one.

Semantic relations are important, of course, because the essential content of texts is not revealed by 'words' alone. To identify the propositions, the attributes and their values, even the phrases that count as 'names' or other lexical atoms in texts, one must appeal to semantic relations. Indeed, all natural-language understanding (NLU) efforts require resources for representing semantic relations and some mechanism to interpret them.

Of course, not all systems that process natural language aspire to natural-language understanding. Historically, for example, information retrieval systems have not attempted to use natural-language processing (NLP) of any sort, let alone processing requiring semantic relations, in analyzing (indexing) free text. Increasingly, however, in the face of degrading performance in large-scale applications, the designers of such systems are attempting to overcome the problem of language variation (e.g., the many ways to express the 'same' idea) and the polysemy of 'words' by using linguistic and semantic resources. Unfortunately, the modest success of computational linguists who have used knowledge bases and declarative representations of lexical-semantic relations to support NLP systems for relatively constrained domains have not been repeated in the context of large-scale information retrieval systems. The reasons are no mystery: no general, comprehensive linguistic knowledge bases exist; no one knows how to build them; even with adequate understanding, no one could afford to build them; the number of 'senses' of words and the

number of relation types is virtually unlimited - each new corpus requires new knowledge.

The interests and needs of information scientists, computational linguists, and NL-applications-oriented computer scientists have been converging. But robust and reliable techniques to support larger-scale linguistic processing have been hard to find. In such a context, the work presented in this volume is remarkable. Dr. Grefenstette has not only developed novel techniques to solve practical problems of large-scale text processing but he has demonstrated effective methods for discovering classes of related lexical items (or 'concepts') and he has proposed a variety of independent techniques for evaluating the results of semantic-relation discovery.

While the results of Dr. Grefenstette's processing are not 'semantic' networks or interpretations of texts, they are arguably sets of related concepts - going beyond a simple notion of synonymy to include pragmatically associated terms. Such association sets reveal implicit sense restrictions and manifest the attributes of more complex semantic frames that would be required for a complete, declarative representation of the target concepts. The key function facilitated by Dr. Grefenstette's process is the discovery of such sets of related terms 'on the fly', as they are expressed in the actual texts in a database or domain. There is no need for 'external' knowledge (such as thesauri) to suggest the structure of information in a database; the process effectively finds empirically-based and pragmatically appropriate term-relation sets for idiosyncratic corpora.

Dr. Grefenstette's approach is distinctive, in part, because it combines both linguistic and information-scientific perspectives. He begins with the observations (that others have made as well) that (1) ''you can begin to know the meaning of a word (or term) by the company it keeps'' and (2) ''words or terms that occur in 'the same contexts' are 'equivalent' ''. He has followed the logical implication: if you find a sensible definition for ''context'' and you keep track of all the co-occurring elements in that context, you will find a 'co-substitution set,' arguably a semantic property revealed through syntactic analysis.

For Dr. Grefenstette, a practical definition of ''context'' is given by a handful of syntactic features. In particular, he takes both syntactic structural relations (such as ''subject of the verb *v*'' or ''modifier of the noun *n*'') as well as general relations (such as ''in the clause *c*'' or ''in the sentence *s*'') to be potential *contexts*. He identifies all the elements (= words or terms) that occur or co-occur in a context and finds statistically prominent patterns of

co-occurrence. Each element, then, can be identified with a vector (or list) of other elements, which are the context markers or attributes of the head element. By calculating 'similarity' scores for each vector, he can find lists of elements that are 'close' to each other, as determined by their shared attributes. These sets are, operationally, the equivalence classes of the words or terms relative to the documents that were analyzed.

Dr. Grefenstette's work is noteworthy in several respects. First, he uses sophisticated (but practical and robust) NLP to derive context sets from free text. This sets his work apart from most other research in information science that has addressed similar concerns. While some scientists are beginning to discover the utility of combining NLP with information-management technology, Dr. Grefenstette has already begun using *varieties* of NLP tools in routine processing. It should be emphasized that Dr. Grefenstette's work has involved serious NLP: his parser is an excellent example of a 'robust' NLP tool practically tuned to its task.

Second, Dr. Grefenstette asks (and answers) *empirical* questions over large quantities of data. This distinguishes his work both from theoretical linguists, who generally have neither the means nor the inclination to work with large volumes of *performance* data, and also from many computational linguists (and artificial-intelligence researchers), whose tools are less robust than Dr. Grefenstette's. In fact, in exploring the thesis presented in this volume, Dr. Grefenstette processed more than a dozen separate databases, representing more than thirty megabytes of text. While multi-megabyte collections are not unusual in information-science applications, they are still rare in AI and linguistic research targeted on *semantic* features of language.

Finally, Dr. Grefenstette focuses his efforts as much on the scientific evaluation of his techniques as on their implementation and raw performance. This is an especially challenging and important concern because there are no well-established techniques for evaluating the results of processes that operate over large volumes of free text. If we are to make scientific progress, however, we must establish new techniques for measuring success or failure. Dr. Grefenstette has embraced this challenge. For example, he evaluates his results from the point of view of intuitive validity by appealing to psychological data. He considers the question of the stability of his techniques by testing whether the results found in increasingly large subcollections of a corpus converge. He assesses the results of (pseudo-)synonym discovery by comparing them to a possible 'gold standard,' two separate, well-known thesauri. He creates 'artificial synonyms' in the data and measures how well his approach can find them. Remarkably, in addition, he implements rival,

alternative techniques for term-equivalence-class discovery and compares the results across techniques. Considered only from the point of view of thoroughness (and cleverness) of evaluation, Dr. Grefenstette's work establishes a high standard.

It is unlikely that any simple approach will be developed that can support the discovery of semantic relations from free text. It is much more likely that we will develop techniques that can assist us in reliably finding limited types of semantic relations. The critical step in creating a general solution will be to combine the results of several restricted and complementary processes in order to achieve a whole that is greater than the sum of its parts. The work presented in this volume is a clear model for such research.

1

INTRODUCTION

The major problem with access to textual information via computers is that of word choice, a problem generated by the basic human ability to express one concept in a variety of ways. This variability raises the question of how the computer can know that words a person uses are related to words found in stored text? Any computer-based system that employs a natural, rather than an artificial, language in its dialogue with human users is faced with this problem.

Some immediately evident problems of language variability are addressed by any computer system that ignores upper and lower case differences, or that allows truncation of suffixes and prefixes. Such character string manipulations are well-understood and ubiquitously implemented, but only scratch at the surface of the problems natural languages cause. The more interesting problem, and one whose solution is not evident, is knowing when two orthographically different terms really signify the same concept. For example, when someone mentions *plasma* should the computer know about *blood*? To answer such a question means leaving the area of simple string manipulation and entering the realm of semantics, the study of meaning.

Feasibility of Solving the Problem by Hand

It may, at first glance, seem sufficient to solve the problem of word variability by simply telling the computer all the important words as well as their relationships within the domain being used. Indeed, this is the approach taken by many natural language interfaces to computer-based systems. Such an approach is feasible (1) when the vocabulary to be used is limited and known ahead of time, and (2) when there exists a person or group of people sufficiently motivated to exhaustively detail the way in which words will be used. Some computer applications which respond to the first criterion are natural language interfaces to particular databases and expert systems

containing complete information about a finely circumscribed domain. As to the second criteria, medical research has generated enough interest to warrant creation of large collections of hand-built and laboriously maintained structured vocabulary such as the MEDLINE thesaurus, usable in computer access of medical journals.

But the overwhelming majority of text available via computer do not satisfy either of the feasibility criteria for manual construction of a network of meaning among the words contained in these texts. The vocabulary is neither known nor limited; and economics does not justify investing the human time needed to begin producing the multifarious networks of meaning over the domains touched upon in the texts.

Why is Discovering Similarity Important?

One reason why discovering similarity is important is that people use a wide variety of terminology, even when talking about the same concepts. Furnas *et al.* (1987) show, consistently across a broad range of domains, that people will use the same term to describe the same object with a probability of less than 20%. Subjects were given examples of common things and asked to give a name for that thing. For example, typists were asked to give names to common editing operations, and the editing operation corresponding to *delete* was called *change*, *remove*, *spell*, *make into*, ... When shown images of common objects, a number of subjects responded *fruit* when shown pictures of nectarines, of pears, and of raisins. We would hope that computer-based systems similarly know when two terms can denote the same concept.

Not only can many different words be used to describe the same concept but each individual word can have a variety of meanings. It is not sufficient to simply have an online version of a common dictionary. A recent study (Krovetz 1991) of a textual database used as a standard testbed for information retrieval systems revealed that words appearing in both the database and in a machine-readable dictionary each had an average of about four different dictionary entries. Computer-based systems must also be able to distinguish which of these dictionary senses the user is employing. This problem of distinguishing meaning appears even harder when it seems (Atkins & Levin 1991) that meanings may not be neat little packages attachable to a word, but rather continua blending from one dictionary sense to another.

These two aspects of word sense variability (many ways of expressing one concept, and many concepts expressed by one word) explain the failure of

simple word-substitution approaches to machine translation, as well as low recall rates in word-based information retrieval systems. In early machine translation experiments (I.B.M. 1959) on a restricted corpus, it was found that, even with a large dictionary of unit-to-unit translations available, the results were very poor. In a large-scale evaluation of a commonly-used information retrieval system (Blair & Maron 1985), it was found that only about 20% of documents relevant to queries were recalled, with analysis showing language variability to be the cause.

Future Need for Automated Discovery Techniques will Increase

With the wide extension of network connections, cheap memory, and powerful machines, more and more textual information in all fields is available on-line than ever before. Idiosyncratic collections of texts can freely and rapidly be amassed, and as quickly dispersed. Wading through such potentially ephemeral information in an efficient and profitable way poses a problem that now stimulates large-scale research (Harman 1993). The automatic recognition of semantic relatedness of disparate, ephemeral, and poorly-established domains becomes important if information is to be extracted from these corpora. Within established, more permanent corpora, if intelligent browsing is to be the paradigm of information retrieval in the future, then a means of browsing from related concept to related concept, rather than from occurrences of stemmed character strings, must be found.

If one cannot afford the luxury of hand-coded semantic knowledge, then recourse must be made to knowledge-poor, domain-independent techniques able to extract semantic information from text. By knowledge-poor, we mean an approach to text that does not necessitate a hand-coded semantic structuring of the domain knowledge before the text can be treated. Such semantic structuring (Mauldin 1991; Jacobs & Rau 1990) might be useful for static domains but manual approaches cannot keep pace with the quantity and variety of text generated.

The future will only see greater demands for knowledge poor techniques, especially if the visions of Vice-President Albert Gore for online information access in every home and business are realized:

> Just as every home in America is now linked to the rest of the country by a driveway that goes to a street, that goes to a highway, that goes to an interstate, we want every home in America to be linked by a national network that will ultimately be . . . invisible to the user in the

> sense that the information will be there at the desktop terminal. (Gore 1992)

Widespread access to unstructured text by many users will only increase the number of idiosyncratic corpora, for which automatic methods will provide the only hope of structuring the information contained in them.

Using Partial Syntax as A Weak Tool

It may seem frustrating to have so much information available electronically, yet not be able to access it with any more than rudimentary tools, since construction of more sophisticated tools would be too costly in terms of human labor. We believe that the computer can be used not only as a repository for this information, but as an active partner in discovering and extracting information from the text it holds, without requiring inordinate human effort.

This book describes a new approach to the problem of computer-based semantics. The approach is that of using weak techniques on a large quantity of texts. By *weak techniques* we mean techniques which require no pre-existing domain-specific knowledge structures nor domain-specific processing, yet which are able to recognize and exploit structural regularities in text. We extract and refine information from text via selective Natural Language Processing using language processing tools that are readily and inexpensively available, rather than using polished techniques that analyze perfectly but which require great investments for small returns.

We use syntax as our motor for generating semantic knowledge, and claim that even partial syntactic analysis provides enough context for judging word similarity of the most characteristic words in a corpus. Completely automated syntactic analysis of common English sentences often remains difficult given the current state of linguistic knowledge. This is so despite the fact that many aspects of language have been studied from a computational point of view over the past fifty years, and that progress on understanding grammars and syntax has been directly translatable in computer language compilers and in database interrogation languages. And though much of the complexity and ambiguity of natural language remain *terra incognita*, some of the more common areas of human syntax are well explored. We use a portion of this well-understood syntactic knowledge as a means of sifting through text, extracting context for words. Comparing the contexts in which words appear allows us to judge similarity between words.

Our research is an attempt to rehabilitate syntax, in a way. It was realized in the early 1960s that, in order to properly perform syntactic analysis, a large amount of semantic and even pragmatic information may sometimes be needed. Since then a tremendous effort has been made to understand what semantics is and which structures are necessary in order to mechanize semantic reasoning. Syntax has been supposed the ultimate beneficiary of this research, but not an active partner. Our research is an attempt to reinject syntax into empirical language studies, as a bootstrapping mechanism for extracting semantic knowledge.

The approach that we take to computer treatment of natural language is hardly classical linguistics, though it is riding the wave of a paradigm shift in computer science. Computer science is still in its infancy, but people realize that the computing paradigm provides tremendous power. It provides more than a tool for implementing some method or theory, but rather provides a different way of thinking of problems. An example is the proof for the 4-color problem (Tymoczko 1990), whose proof would have been impossible without the computer paradigm. Our approach to language, although grounded in statistics rather than proofs, is also based on sifting through massive amounts of data using simple techniques. As in gathering wheat, if we are not concerned with catching every grain, we can rapidly harvest large amounts of useful material with simple methods.

The structure of this book is the following. After an historical overview of computer approaches to semantics in Chapter 2, Chapter 3 provides a description of a robust domain-independent partial parser for English that yields local syntactic contexts of words. That chapter explains a method we devised to use these contexts to generate dependent corpus-dependent similarity lists. Chapter 4 provides evaluations of the results of our methods applied to a number of corpora. We demonstrate that the similarities extracted by this method correspond to human similarity judgments as described in psychological literature and as captured in manually created thesauri. In Section 4.3, in particular, we demonstrate that the overlap with manual thesauri using this technique is greater than that obtained by traditional text windowing techniques. In Sections 4.2 and 4.3 we present evaluation methods that we created that are applicable to any corpus-based meaning extraction techniques. In Chapter 5, we show applications of our similarity discovery techniques to information retrieval, thesaurus enrichment, and automatic thesaurus construction. In the concluding Chapter 6, we discuss further axes of development of our discovery techniques, and show how such techniques can be applied to multi-word phrases. Appendices and a Bibliography follow.

2

SEMANTIC EXTRACTION

2.1 HISTORICAL OVERVIEW

The real impetus for development of computer-based techniques for dealing with semantics was the realization that simple word substitution was inadequate (I.B.M. 1959) for machine translation. It was realized that context influenced word meaning and that each word's context would have to be taken into account. The early information retrieval community was also interested in semantics classification, but for a different reason. Information retrieval, since it arose as a science from library science, which itself had a long history of classification, was interested in implementing an online version of human classification schemes and was concerned in a subsidiary manner with automating this classification process.

This research developed then in two directions. One branch took a long-range approach to the problem and studied which structures and mechanisms would be needed in order to perform high-quality language understanding. This branch subdivided into the fields of Cognitive Psychology, subbranches of Artificial Intelligence, and Computational Linguistics. The other major branch attempted to find short-term solutions to language variability problems by exploiting available knowledge sources with available techniques, taking a more direct engineering approach. This approach is best illustrated by work done in the field of Information Retrieval.

We first examine the more ambitious approaches undertaken by researchers in the field of Computational Linguistics and Artificial Intelligence. Although these approaches hold promise, their application to problems of extracting information from unrestricted text is premature. Next we examine the more

```
bachelor
  1. (human)(male)           "a man who has never been married"
  2. (human)(male)(young)    "a knight serving under the
                              standard  of another knight"
  3. (human)                 "who has the first or lowest
                              academic degree"
  4. (animal)(male)(young)   "a seal without a mate
                              during breeding time"
```

Figure 2.1 Katz and Fodor's semantic markers for the word 'bachelor.'

engineering-like approaches to semantics. The central thread running through this research is a reliance on textual data rather than on cognitive theory in trying to extract semantic classes. We examine extraction techniques from structured and then from unstructured texts. In the final section of this chapter, we discuss the contemporary approaches most related to our technique.

2.2 COGNITIVE SCIENCE APPROACHES

2.2.1 Semantic Markers

Inspired by techniques proven to be successful in the syntactic domain, Katz & Fodor (1963) initiated a tradition of semantics as manipulation of semantic markers attached to lexical items. Just as lexical items could be adorned with syntactic markers, which could then be used to analyze the structure of sentences (Chomsky 1957), this semantic theory posited markers such as *(human), (male), (animal), (young)* that would be used to build the possible senses of every word. For example, the four senses of the English word *bachelor* would be described by appropriate semantic markers as shown in Figure 2.1.

To decide the meaning of any word in a given sentence, a postulated body of rules would describe how these markers could permissibly interact in non-anomalous sentences. Choosing the proper sense of a word in this theory would follow mechanisms akin to choosing its part of speech. For example, the phrase *the bachelor's fur* would select the fourth meaning of *bachelor* since *fur* somehow activates the *(animal)* marker.

Such an approach was attacked as untenable for a variety of reasons. Early on, Bolinger (1965) noted that although phonetics and syntax possess a limited number of contrasts between the units they treat and that, therefore, the number of markers needed can be limited, the same cannot be said of semantics. Semantic rules must distinguish among a potentially unlimited number of contrasts. In addition, semantic markers ultimately suffer from the same problems as the original lexical units: redundancy, multiple meaning, circularity, and world knowledge necessary to disambiguate them. As just one example of this last problem, he considered whether *(early)* and *(young)*, two markers suggested by Katz, were related or unrelated markers, something which must be known if semantic operations such as comparison or analogy were to be possible. Later, Lewis (1972) complained that the use of a large number of such markers results in nothing less than the creation of a new language, that he called ‘‘Markerese,’’ into which a sentence is translated and whose elements must once again be interpreted. Eco (1984) argues that if markers cannot be interpreted then we, or a machine, can never know what something is. For example, how can a machine determine when something possesses the primitive marker *(animal)*? One the other hand, if markers must be interpreted, then one cannot limit their number. For example, an *(animal)* must be defined in terms of other markers, which in turn are defined in terms of others, and so on. G. E. Barton *et al.* (1987) argues that the explosion of markers makes for a system whose verification is computationally intractable, since the number of cases that the rules must distinguish between grows exponentially in the number of markers.

Despite such criticisms, the computational attractiveness of being able to implement semantics using a (hopefully) limited number of markers still exerts a great influence in the computer science community, especially among researchers in Artificial Intelligence. Though few researchers still believe that such semantic marking techniques will be applicable to the scope and variety of unrestricted text to which syntactic markers can successfully be applied, it is hoped that in situations where the semantic domain to be treated is small and well-delimited, such techniques can be useful.

2.2.2 Slots, Frames and Scripts

The initial idea of binary markers was expanded to slots and frames (Minsky 1975) by the Artificial Intelligence (AI) community. Instead of simply possessing a marker, each lexical entity could contain slots in which were found either a value, a pointer to a default value, or a procedure that supplies

```
"policeman" --> ((FOLK SOURCE)
                 ((((NOTGOOD MAN) OBJE) PICK) (SUBJ MAN)))
```

Figure 2.2 Wilks' semantic formula for 'policeman.'

the missing value. Using such a slot-based system augmented by a planner, Winograd was able to create a sophisticated interactive language understanding system that responded to user queries and to user commands by altering the state of a simple blocks world (Winograd 1973). Winograd (1972) argued that such a system is necessary to perform language understanding at a level sufficient for quality machine translation. For example, when passing from English to another language, understanding which of many possible referents of neutral anaphoric pronouns is the correct choice (e.g., knowing what *they* refers to) is necessary for choosing the proper gender for target pronouns. Knowing the semantic constraints between lexical entities could solve certain of these cases.

In the same tradition, Wilks (1975) proposed an *Intelligent Analyzer and Understander of English*, in which words were described by ''70 primitive semantic units used to express the semantic entities, states, qualities, and actions about which humans speak and write'' [p. 266]. Some examples of these primitives are `MAN`, `STUFF`, `SIGN`, `THING`, `CAUSE`, `PICK`, `GOOD`, `GOAL`, `SUBJ`, `THRU`, `LOCA`, `SOURCE`, `IN` , and `POSS`. Primitives could be composed in two-element subformulas such as `(FLOW STUFF)` for liquidity. Each lexical item possessed a semantic formula composed of sub-formulas. The extended formula for *policeman* is given in Figure 2.2.

This formula can be read as ''[the policeman is a] person who selects bad persons out of the body of people'' [p.267]. Verbs formulas imposed certain classes of primitives on their subjects and objects and a set of semantic templates imposed certain possible orderings of words in a sentence. Wilks' work was one of the closest implementations of the original ideas suggested by the Katz and Fodor model.

He demonstrated that such a system was powerful enough to do translation, disambiguating word senses in certain sentences formed from the 600-word vocabulary he had implemented. Scaling up was evidently a problem, as can be seen by the example given above. Obviously, the function of arresting a criminal is only one of the myriad functions performed by a policeman. Each

conceivable function would have to be encoded in the way illustrated in order to treat general text.

A more elaborate representation of semantics, one that combined Katz's and Fodor's theories with the more traditional linguistic tradition of case grammars (Fillmore 1968), was developed in Schank (1975). In Schank's representation of lexical semantics, that he calls Conceptual Dependencies (CD), verbs were described by semantic primitives such as `ATRANS`, `PTRANS`, `MTRANS` (for abstract, physical, and mental transfers), `PROPEL`, `MOVE`, `GRASP`, `INGEST`, `EXPEL`, `SPEAK`, `MBUILD`, etc. These primitives were augmented graphically by diverse arrows indicating case relations with different nouns that were objects, instrumentals, locatives, datives, or subjects of the verbs.

As a means of using this representation to choose word senses, scripts or stereotyped sequences of actions (Schank & Abelson 1977), and later Memory Organization Packets (Schank 1982), were developed to lay over a given text in order to select permissible meanings, much as with Wilks' templates. An example of such a script is the sequence of CDs to be found in a typical American restaurant setting. Lehnert (1978) presents a detailed description of work inspired by this research, that unfortunately suffers from the same scaling problems as Wilks' system.

2.2.3 Text Skimming

More recent contributions to this Schankian tradition are FRUMP (DeJong 1982) and FERRET (Mauldin 1991). Both involve skimming over text, filling up CD structures when possible, and using filled CDs to determine which script applies to the text. Determining which script should be activated allows the system to establish a domain that, it is hoped, will distinguish word senses, much as establishing that one is talking about animals permits the proper choice among senses of *bachelor* in Katz and Fodor's original project.

FERRET used an online dictionary to look up alternative synonyms for unknown words. For example, if the system came across the word *venture*, for which it had no CD, (1) it looked up the definition of *venture*; (2) finding **venture: to proceed despite danger**; it then tried to find a CD for one of the words in that definition; (3) this failing, it proceeded recursively until, in the definition of *proceed*, it found **proceed: to move along a course :: advance**; (4) in this definition FERRET came across the word *advance* for which it has a CD to use as a model for *venture*. Mauldin claims that FERRET

claims produces significant improvement in both *recall* and *precision*, versus standard boolean keyword search, in an experiment with 1065 astronomy texts and 40 queries generated by graduate students. This improvement should not be surprising, considering that a great quantity of information about astronomy was stored in FERRET's handcrafted CD scripts. This additional information allows for the domain-directed inference and abstraction which are obviously missed by unexpanded keyword search. The need of intensive domain-dependent hand-coding limits the extensibility of this system.

Jacobs & Zernik (1988) also proposed a method of acquiring word senses of unknown words, given a context of known words for which CD-like structures existed. The unknown word filled the CD-slots of known words in the sentence; then previously encountered CDs from the corpus were compared to find the best match, with the unknown word being hypothesized as similar to that matching word. A demonstration of the Jacobs and Zernick method on a few lines of text was provided. Their technique most likely corresponds to the way humans acquire new vocabulary, but its implementation would require that most of the scripts and CDs necessary to understand a certain domain already exist. In certain message understanding tasks and in domain filtering tasks (Jacobs & Rau 1990), such a collection of CDs have been built but always requiring a massive investment for a narrowly circumscribed subdomain.

2.2.4 Augmented Transition Networks

Another approach to sense disambiguation can be seen in the Augmented Transition Network (ATN) approach (Woods 1970). This technique uses a finite state automaton formalism whose power is augmented by registers and procedures that can be invoked at each node.

This approach has proved useful when the structure of the phrases to be used is known ahead of time and when the domain of speech to which the system is applied is tightly delimited. Examples of successful implementations of this technique are the LUNAR program (Woods *et al.* 1972), that could answer questions about moon rock data. ATN grammars are equivalent in power to a programming language and can be complex to program, making robustness arduous to achieved.

A similar technique that introduces semantic and lexical items in the lowest level of syntactic analysis is that of semantic grammars (Hendrix 1977; Hendrix & Lewis 1981). In this approach, a context-free grammar includes

```
S              -> what is SHIP-PROPERTY of SHIP ?
SHIP-PROPERTY -> the SHIP-PROP | SHIP-PROP
SHIP-PROP      -> speed | length | draft | beam | type
SHIP           -> SHIP-NAME | the fastest SHIP2 | ...
SHIP-NAME      -> Kennedy | Kitty Hawk | Constellation | ...
...
```

Figure 2.3 A small semantic grammar for a naval data base.

semantic classes that resolve to lexical items. In the example given in Figure 2.3 that was drawn from a system that could answer questions about a naval fleet, semantic properties are represented as upper-case words, terminal strings are in lower case, and the grammar describes the structure of recognized input queries.

These techniques have been successfully integrated into commercial natural language front-ends for traditional relational databases (INTELLECT 1982; Harvey 1988). They can be used with databases, since each term appearing in a query on a database appears in a limited number of fields that determine its semantic properties.

For example, Figure 2.4 shows the transformation that a user query in natural language undergoes in a natural language front-end. The grammar recognizes that `WHO WORKS IN` is a request on the name field, that `FRENCH` is a synonym for a the value `FRANCE` in the `COUNTRY` field, and that `SALES` and `ACCOUNTING` are department names.

2.2.5 Semantic Nets

The above approaches to computational semantics consider lexical units in isolation from one another. Either each sense of a lexical item possesses a certain number of markers or slots, or each lexical item is explicitly included in some part of the grammar, in a position that defines its sense.

A different approach proposed in Quillian (1968) does not explicitly define the senses of a word but rather links all the lexical items into a large net. This avoids the problem of defining a set of primitives, since the primitives of meaning are the lexical items themselves. Quillian took a knowledge-poor approach to this problem since he did not suggest adding any new

USER INPUT:

```
"WHO WORKS IN THE FRENCH SALES AND ACCOUNTING DEPTS?"
```

DATABASE REFORMULATION:

```
SELECT NAME
FROM EMP DEPT
WHERE (DEPT.COUNTRY="FRANCE" AND
      (DEPT.DNAME="SALES" OR DEPT.DNAME="ACCOUNTING")
       AND DEPT.DEPT# = EMP.DEPT# )
```

Figure 2.4 A natural language front-end for database querying.

```
(cl-define-concept 'tradable-object
                        '(and classic-thing
                          (all users human) ))

(cl-define-concept 'game-object
                        '(prim inanimate game-object))

(cl-create-ind 'marbles '(and
                            game-object
                            (and tradable-object
                               (all users children)) ))
```

Figure 2.5 A knowledge representation language for describing concepts.

explicit semantic information to the lexical items appearing in a standard dictionary. His approach and experimentation are discussed below (p.19). His ideas nonetheless inspired AI researchers who were devoted to knowledge intensive approaches to language. They created methods of structuring domain knowledge into semantic nets (Brachman & Schmolze 1985; Sowa 1990) in which what corresponds to slots in the frames approach above are replaced by labeled links to other lexical items.

For example, the knowledge representation language CLASSIC allows one to define concepts such as those in Figure 2.5.

The above declarations define concepts such as `game-object`, that is linked to a class of `inanimate` objects, `tradable-object`, that can

have `users` only from a class of `humans`. Moreover, they tie the lexical object *marbles* to both concepts, restricting the class of `users` of *marbles* (`tradable-objects`) to the subclass `children`.

The most ambitious project using semantic nets is the CYC project (Lenat *et al.* 1986). The CYC project was initially a large-scale attempt at using semantic net methodology to supply the background knowledge, be it physics knowledge, social knowledge, psychological knowledge, or whatever, that would be needed to understand written information sources such as newspapers, advertisements, or encyclopedias. In a talk (1990) at Carnegie Mellon University, Lenat qualified his research by stating that he no longer proposes to produce something that is ‘‘use-independent.’’ He stated that he and his co-workers found that the encyclopedias from which they had hoped to draw ‘‘common sense knowledge’’ actually contain what might be considered the complement of common sense. As an example, he cited the *Encyclopedia Britannica* article on alchemy. Common knowledge, he said, leads us to think about alchemists changing lead into gold, yet in the alchemy article all sorts of famous alchemists and their works are described but not once is lead or gold mentioned.

2.2.6 Summary of AI approaches

The commonality of all the above approaches (semantic markers, slot-based systems, CDs, ATNs, semantic nets) is the presupposition of encoded world knowledge in the structures attached to each lexical item. When language variability is encountered in real text, this world knowledge hopefully allows the elimination of ambiguities and the recognition of similar concepts. In all of the experimental and commercial systems listed above, this semantic information is hand-coded into the system.

Though such structures are useful when the domain is well-defined, they are expensive to construct, verify and maintain. Many of these systems cited in research were never intended to be practically exploitable, but instead were meant as models for investigating certain problem areas of human understanding and reasoning. Concentrating on these theoretical issues, their creators rarely demonstrated the feasibility of implementing their systems or of attacking the problems posed by actual, unrestricted text. Lenat *et al.* (1991) characterized these systems in the following way:

> RLL, KRL, etc., ... were ''implemented serially, at the demo level,'' which means that each of the documented features worked, at least once, in some demo, though few of them worked simultaneously, or repeatedly, or on a wide range of problems. This ''sort of'' level of implementation was not limited to Stanford, or the 80's, of course; one could point one's finger at SHRDLU, HACKER, ... well, you get the idea, and as I've already remarked, I was as guilty as anyone of this sloppiness.
>
> Living in this never-never land of semi-implemented systems has certain advantages (it lets you expound, and cling to, elegant theories of knowledge and intelligent actions). This is what Aristotle did with physics; it's what we're *still* able to do with religion. Things get ever so much uglier if you let data intrude. Other than carefully selected examples, of course.

Though the practical applications of such hand-coded systems are limited, their principles have found use in creating natural language interfaces to traditional databases. The advantages of traditional databases over ordinary text collections is that individual terms are semantically specified by the field in which they are found. But this limited success is not the only justification for continuing research in this approach to semantics. Such fundamental research in understanding how semantics can be codified and reasoned with must continue if intelligent systems are one day to understand and correctly translate unrestricted text.

Even now many bodies of text exist, and more are being created every day, for which we cannot spare the expense or time to manually form semantic structures. Though current systems can readily index text and retrieve strings, such text will be poorly exploited unless some means is found of extracting the information in it by automating a large portion of semantic structure encoding.

2.2.7 Problems with AI approaches

The major disadvantage of hand-coded semantic structuring is scalability. Although there are fewer than one hundred thousand different words in a standard English lexicon, coding a semantic structure is much more tedious than repeating one hundred thousand encoding operations. Since the structures cover not only each individual lexical item, but also all the various senses of each word, as well as all the possible relations between words, the problem quickly becomes much more complex.

The problems with applying manual encoding approaches to the problem of extraction of information from raw text are given below.

Cost: There are costs of initial coding, of coherence checking, and of maintenance after modifications, and costs derivable from a host of software engineering concerns.

Domain dependence: A semantic structure developed for one domain would not necessarily be applicable to another. For example, *sugar* would have very different semantic relations in a medical domain and in a commodities exchange domain.

Rigidity: Even within well-established domains, new subdomains (cf. AIDS treatment in Medicine) spring up and become important. Can hand-coded systems keep up with new discoveries and new relations with an acceptable latency?

Reliability of Coding: Since producing a knowledge structure describing a domain means imposing a certain perspective on it, one may hope that a single programmer can maintain a relative coherence in his or her coding. But in any larger project, a knowledge structure will be the result of many programmers each with their own point of view and coding styles. The question of how the structure's reliability and coherence can be maintained must be answered.

2.3 RECYCLING APPROACHES

The AI techniques for treating natural language have been labeled above as being knowledge-rich since they require a large investment of entering domain-specific knowledge into computer-based structures before they can be applied to the treatment of text. The cost of creating and maintaining this knowledge has been of course recognized and has led some researchers to explore possible ways to accelerate or to automatize its acquisition. This perspective has motivated a number of knowledge-poor approaches to bootstrapping domain-specific semantics from existing sources.

One knowledge-poor approach to computer extraction of semantics is to exploit text appearing in documents whose semantic structure is known. Such a technique can be seen as salvaging or recycling specific human judgments as to how words are related. Such documents were constructed by humans with

the precise view of explaining to other humans the semantics of a domain. When explaining things to humans many different levels of understanding (grammatical, syntactic, experiential, historical, socio-cultural, encyclopedic) may be interleaved in the explanation. The task of the knowledge-poor technique is to recognize patterns that can be mechanically exploited without reference to these deeper levels of understanding.

2.3.1 Using Dictionaries

Sparck Jones's pioneering research, done in the early 1960s (Sparck Jones 1986) attempted to use dictionary definition senses to define semantic heads, just as *Roget's Thesaurus* classifies words under 1000 headings. Many words in dictionaries possess a certain number of numbered senses. In this research, each dictionary sense was manually reduced to the principal nouns appearing in the definition. The reduction was made by examining the sample use phrase that accompanied each dictionary definition in the *Oxford English Dictionary* and by substituting all the nouns appearing in the definition for the head word in that phrase. For example, one sense of *task* is

> **Task** 3 a) A piece of work which has to be done; something that one has to do (usually involving labour or difficulty); a matter of difficulty. ''He had taken upon himself a task beyond the ordinary strength of man.''

From this sense of *task*, Sparck Jones would create the **row**: *task - labour - work*. In her research, such a row corresponded to some concept for which *task*, *labour* and *work* could be synonyms.[1] Rows were clustered by considering each element in the row as an attribute, and then clustering rows having similar attributes. Limited experiments were performed due to computational restrictions at the time. All evaluations were based on intuitive judgments of correctness of the clusters formed.

More recently, Plate (Wilks *et al.* 1989) performed similar experiments with dictionary senses from the *Longman Dictionary* (Proctor 1978). This dictionary was specifically designed to use a restricted set of primitive words, e.g., *girl*, *woman*, *ceremony*, *nation*, *relation*, *occasion*, *king*, ..., in its definitions. These 2200 primitives were clustered using dictionary sense co-occurrence. Plate writes, ''The exact nature of the relationships that the co-occurrence data

[1] Later, in Section 5.3, we show how such `rows` can be derived from raw text.

reflects is unclear.'' But the results seemed good. A psychological validation of the clustered results was performed. In a typical experiment, a human judge would be presented with a group of 20 primitive words and asked to rate the relatedness of all 190 pairs of words. Human subjects consistently found that the groups of words that had been clustered together did seem more related than words not clustered.

Quillian's original research on semantic nets (Quillian 1968) also aimed to exploit existing dictionaries. He described the semantics of a word sense as being a plane containing all the nodes in that word sense's definition. Each plane has a certain number of links into it, corresponding to the other definitions that use that word sense; it also has a certain number of links going out, which correspond to the word senses comprising the definition. This model uses words themselves, as arranged in a dictionary, as the semantic model. The semantic relation between words is described by describing the links between their two planes and any intermediate links between them.

Quillian ran experiments to provide intuitive evidence for this model. Definitions for 850 words from a basic English dictionary were entered into a computer. Multiple senses of words, such as ''plant: a living organism" and ''plant: a factory", were distinguished by manually adding an index, e.g., `PLANT1, PLANT2`, to every ambiguous word. Here is a sample definition: `COMFORT3 : GIVE5 STRENGTH2 OR MAKE2 LESS2 SAD.` Figure 2.6 describes one of Quillian's experiments.

One of the problems with this approach is that manual disambiguation of word senses within dictionary entries must be made beforehand; e.g., someone has to decide when a `MAKE` that appears in a dictionary entry is a specific sense, say `MAKE2`. Another problem is that the search of intersecting nodes is very costly in time and space, $O(N^3)$ where N is the number of words in the dictionary. Quillian was unable to test his system on more than 20 definitions at a time and no method of evaluation of the resulting intersection was proposed.

In the past ten years Machine Readable Dictionaries have become more available and the *Longman's Dictionary of Common English*, which uses a reduced vocabulary in its definitions, has been exploited for semantic extraction. Vossen *et al.* (1989) used regular structures of the definitions in this dictionary to extract a limited number of semantic hierarchies. Since this dictionary was constructed as a learning dictionary, many definitions are of a predictable form, e.g., ''(word): a (word2) that ..." as in the definition given below.

```
Experiment:
  User gives two words,  such as CRY and COMFORT.
  Each word definition is considered as a collection
  of nodes,  with each node corresponding to one word in
  the definition.
  All the nodes leading from the definitions of each user
  word are explored until the two searches meet.
  Their meeting point is displayed:
  Examples,
  Compare: CRY, COMFORT
  Result: 1st Intersect: SAD
         (1) CRY2 IS AMONG OTHER THINGS TO MAKE A SAD SOUND
         (2) TO COMFORT3 CAN BE TO MAKE2 SOMETHING LESS SAD

  Compare: EARTH, LIVE
  Result: 1st Intersect: ANIMAL
         (1) EARTH IS A PLANET OF7 ANIMAL
         (2) TO LIVE IS TO HAVE EXISTENCE AS7 ANIMAL
```

Figure 2.6 One of Quillian's semantic net experiments.

```
sf(crook1,
   [[arcs,
     [[supertype, criminal1]]],
    [node0,
     [[it1, steal1,
         valuables1]]]])
```

Figure 2.7 Knowledge structure automatically extracted from a dictionary

> anaesthetist: a doctor who gives an anaesthetic to a patient ...

This definition shows that an *anaesthetist* is type of *doctor*. Once such structures have been inventoried, Fass (Wilks *et al.* 1989) believes that they can be used to semi-automatically build up sense-frames for words, as in the structure, given in Figure 2.7, that might be extracted for the word *crook*.

This approach raises a number of problems, since the patterns recognized are those of a restricted dialect of English, i.e., dictionary entries, and, as is the case with Quillian's original research, there remains the problem of disambiguating

```
cough-ronchi
bleeding-hemataroses
egotism-greediness
stupor-apathetic
```

Figure 2.8 Words often co-occurring in labeled fields of a medical thesaurus.

senses of words used within definitions, e.g., *steal1* or *steal2*. Guthrie *et al.* (1990) have presented one method, however, of using the semantic codes present in *Longman's* to disambiguate genus terms.

2.3.2 Compendia and Thesauri

As another example of using a structured information source, Blois (1984) extracted lists of related words from an an online version of *Current Medical Information and Technology* (CMIT), which contained 3262 descriptions of different diseases.

In the CMIT, each disease description is divided into 11 fields, such as cause, body part affected, symptoms, laboratory test, etc. An entropy measure was defined giving lower entropy to words appearing many times in few fields. This measure divided the vocabulary into a first third of medical terms, a bottom third of non-medical terms and a middle third of medical/non-medical terms.

One thousand low entropy terms were selected from the symptoms and signs fields. The co-occurrence of these words anywhere in the same disease description was counted. An association measure that compared the observed term co-occurrence rates with those expected from each terms single-term frequency was defined and computed. Words such as those given in Figure 2.8 appeared together much more often than chance would dictate.

This research seemed to succeed in bringing together words which share semantic properties by simple statistical means, though no evaluation of correctness was performed. But even the intuitive results are biased by the fact that the structure of the CMIT permitted the easy recognition of words describing symptoms and signs (special fields), words which already share the same semantic roles of providing clues to illness.

In Peretz Shoval's PhD dissertation (Stengel 1981), using an online thesaurus was proposed as the first step in building up a semantic net. The idea was to exploit the explicit relations given in a manually-constructed thesaurus, as well as the implicit relations between words composing a noun phrase; for example, *information* and *science* are more general terms than *information science*. All of the terms found in the thesaurus were to be stored in manually augmented hierarchical semantic lattices. When an end-user supplied key words, this lattice would be traversed to find the most specific terms that were descendants of the original terms. An evaluation of users' responses to a prototype of this system showed that they preferred having pertinent terms suggested from the machine in this way to manually searching an on-line thesaurus.

2.3.3 Problems with Using Structured Sources

This idea of using a knowledge-poor method to exploit semantically rich and well-structured textual sources such as on-line dictionaries or thesauri offers much promise for extracting relations among words of the general vocabulary. In addition to the reserach presented here, much work is currently being conducted on acquiring semantic information from machine readable dictionaries (Montemagni 1993; Calzolari 1991; Carroll & Briscoe 1989). Decades of man-years have been expended in creating these dictionaries providing a rich semantic capital, yet these resources also suffer from the following drawbacks:

Closedness: Dictionaries are long-term investments which are rarely updated.

Bias: A team of human lexicographers used their own bias to decide what to include in and exclude from the dictionary. Although this is the value of a manually created dictionary, it limits the applications of the dictionary.

Incompleteness: General-purpose dictionaries do not contain domain-specific terms. Not only are proper nouns missing but also common words in a domain, as well as domain-specific uses of common English words.

Definition Variability: The formats of the definitions, intended for humans, are variable. They sometimes resemble a semantic marker list, sometimes being contrasting with some other object, sometimes describing a real-world situation (Allen 1992).

Multiple Senses: Some mechanism must still be provided for deciding which of the given senses apply. The overlap between senses in diverse dictionaries is problematic (Atkins & Levin 1991).

Esoterism: Rarer words and rarer senses are included among common ones without indication of their prevalence. Some dictionaries such as the OED provide the oldest historical sense of a word first, while others such as Webster's give the most common meanings first. Yet a measure of commonness or word frequency in general English is never provided.

The same criticisms of closedness and incompleteness can be leveled against any technique using any fixed non-renewable text. Though the resource may be exploited, the question remains: Once the text has been exploited, how is one to continue? We are interested here in this work of exploring techniques that are applicable to renewable resources.

2.4 KNOWLEDGE-POOR APPROACHES

Starting from Section 2.2 we saw how a machine-intelligence approach to the language variability problem necessitates a large quantity of knowledge, and how this need stimulated research on extracting such information from structured sources. But we also discussed how knowledge-intensive approaches suffered problems of scale and how structured sources are limited in number. At the same time as these research paths were being followed, a number of researchers took a different route, examining what information can be extracted from unstructured text, using little or no knowledge. Most of this early knowledge-poor semantic extraction work was based on the statistics of co-occurrence of words within the same window of text, where a window can be a certain number of words, sentences, paragraphs, or an entire document.

2.4.1 Document Co-occurrence

Antonyms, Synonyms and Semantic Axes

In an early demonstration of the power of using simple counting techniques as a key for uncovering semantics, Lewis *et al.* (1967) showed that, using Chemistry Ph.D. titles as documents, synonyms tend never to occur together, but often tend to co-occur with the same set of other title words. Justeson &

Katz (1991) have demonstrated that, over large corpora, antonymous adjectives tend to co-occur in the same sentences much more often than frequency-based probability would indicate.

These results are interesting because they indicate that the simple counting of words and the other strings occurring with them may indicate which words belong to classes of similar meanings. One of the aspects of language variability is that many different words can be used to describe the same concept, and here we have indications that an automatic means of discovering the words associated with a concept is possible. If each concept is visualized as an axis in the space of all meaning, then one hypothesis for developing this automated discovery is that when one concept is being discussed in two different texts there will be a wide overlap of words being used to describe it. This hypothesis is at the basis of almost all document-document comparison schemes in the information retrieval community (Treu 1968; Salton & McGill 1983; Croft & Thompson 1987). It has also been used in term-term comparison experiments (Srinivasan 1992).

Deerwester *et al.* (1990) used document co-occurrence to build up a data matrix where each row represents a word and each column represents a document from some corpus of documents. The entry in each matrix position corresponds to the presence of that word in each document. They then used singular value decomposition to reduce the matrix to its principal axes. This has the effect of reducing the space described by all the words to a smaller space of semantic axes, reducing the problem from thousands of dimensions to hundreds. Each word can then be thought of as a point in this reduced space, specified by its value along each dimension's axis. By considering the distance between words in this space, semantically related words appear closer together. The composition of all the words appearing in the query on the corpus also defines a point in this reduced space, and documents found near that point are chosen in response to a query. Deerwester et al. have shown that this semantic space reduction can improve information retrieval. This technique suffers from the drawbacks of (1) computational complexity since matrix reduction is $O(N^3)$ where N is the smaller of the number of terms and the number of documents, and (2) attacking only one part of the language variability problem, that of many terms concerning the same concept. Indeed, the other aspect of language variability, that one word can mean many things, introduces noise into the calculations of the semantic axes. Schutze (1992) uses a related technique called canonical discriminant analysis to create semantic axes, using co-occurrence of terms within windows of 1000 characters, which suffers from the same computational complexity.

Neural Nets

Document co-occurrence has also been used in a neural nets approach to information retrieval (Kwok 1991). In this experiment, a two-way neural net is built having three layers, a query layer, a term layer, and a document layer. In a learning phase, certain relevant documents supply terms which are used to reinforce the weighting between the layers. In this way entering a query excites term nodes not necessarily present in the original query but known to exist in relevant documents. This technique again attempts to reduce the semantic space to a limited number of axes represented by the hidden layer of nodes. Here, the number of nodes in the intermediate layer of the neural nets corresponds to the number of orthogonal factors retained in the singular value decomposition technique described above, which itself corresponds to the number of semantic axes retained in the space reduction. One of the principal drawbacks of this technique is that it necessitates manually created query-document relevance judgments before any weighting can begin, thus limiting the technique to a restricted set of domains.

2.4.2 Textual Windows

A more classic approach to knowledge-poor semantic extraction using co-occurrence is to use a small window, e.g., four or five words, to extract the words commonly surrounding each word. This technique is easy to implement since it requires no lexical information whatsoever; usually a stoplist of a hundred or so 'empty words,' such as articles, prepositions, etc., are eliminated from consideration as bearing little information. The context extracted around each word is used in two ways: to calculate which words appear together often, and to see which words share the same contexts.

Phillips (1985) used this information to construct clusters which he called 'lexical networks.' The goal was to link together portions of each document which shared a significant number of networks, in order to reveal the global structure of the document. Clusters produced within individual chapters were compared manually to section headings. For example the cluster {*flow, involve, mass, units, control, referred, length*} was one extracted cluster which Phillips found in, and considered related to, the Section Heading ''Units for Mass, Length, Time and Force" in the tested Thermodynamics text. These lexical networks were also used to link chapters in the following way. If two networks from different chapters shared at least two terms, then a 'link' was created between the chapters. If two chapters shared two or more links

honorary	doctor
doctors	dentists
doctors	nurses
doctors	treating
examined	doctor
doctors	treat
doctor	bills
doctor	visits
doctors	hospitals
nurses	doctors

Figure 2.9 Words co-occurring often with 'doctor' within a window of five words in a large corpus of newspaper text.

then a path between them was created. Phillips claims that the examination of the paths developed between chapters thus corresponds to the textbook's structural division described by the textbook authors in their prefaces. Hence, these automatic techniques reveal the macro-structure of the text according to the author.

Church & Hanks (1990) use textual windows to calculate the mutual information between a pair of words. They employ an information theoretic definition of mutual information which compares the probability of observing two words together to that of observing each word independently. Words having high mutual information over a corpus are usually semantically related. For example, using this mutual information criterion and a window of five words over a corpus of 15 million words from a news service, words given in Figure 2.9 were highly associated with *doctor*. This technique highlights common noun phrases such as *doctor bills, doctor visits*, common conjuncts such as *doctors and nurses, doctors and dentists*, and common noun-verb pairs such as *doctor treats, doctor examines*. Such results suggest that recognizing these types of lexical-semantic units may provide clear semantic relations.

In Brown *et al.* (1992), a much larger window of 1000 words excluding the 5 words directly around each word was used to measure mutual information. A clustering technique, that was aimed at maximizing the average mutual information within clusters, was then iteratively applied to pairs of clusters to create a specified number of general semantic classes such as {*tie, jacket, suit*}, {*morning, noon, evening, night, nights, midnight, bed*}, or {*problems, problem, solution, solve, analyzed, solved, solving*}, from a corpus of 365 million words from a variety of sources. This knowledge-poor technique,

based solely on counting strings, provides interesting results, though it is computationally expensive, $O(N^3)$ with a large coefficient and where N is the number of distinct strings.

2.4.3 Problems with Co-occurrence methods

These knowledge-poor techniques of using numbers of occurrence or co-occurrence of words within a document or within a window are certainly applicable to corpora from any domain. But document co-occurrence, the most commonly used statistic, suffers from at least three problems:

Granularity: Every word in the document is considered potentially related to every other word, no matter what the distance between them. For example, words from the beginning and from the end of a document will be brought together as a data pair by such technique, though there may be no longer any connection between the subjects discussed at those two points. When smaller windows are used similar effects are still seen, as can be seen by this very sentence where the word *similar* falls within a 5 word window around the first appearance of *window* even though they belong to different noun phrases.

Co-occurrence: For most semantic grouping techniques using document co-occurrence two words are only seen as similar if they physically appear in the same document a certain number of times. As a trivial counter-example, consider the words $tumor$ and $tumour$. These words certainly share the same contexts, but would never appear in the same document, at least not with a frequency to be recognized by any document co-occurrence method. In general different words used to describe similar concepts may never be used in the same document, and are thus missed by these methods.

Unbounded Growth: Since similarity based measures are generally $O(n^2)$ or $O(n^3)$, where $n = max(NbObjects, NbAttributes)$, using 'presence in the same document' as an attribute means restricting oneself to small corpora of a few thousand documents. As more documents are added a level of intractability is quickly reached, as seen in work using this attribute (Deerwester *et al.* 1990).

Using a window of words restricts the granularity of co-occurrence but seems to be merely palliative to correctly recognizing phrases, useful in times before

robust noun-phrase extractors (Grefenstette 1983; Evans *et al.* 1991b) or syntactic analyzers (Hindle 1989) were available.

2.4.4 Salient Noun Phrases

The language variability problem of one word having many nuances of meaning is tempered when the word appears as part of a longer term. For example, humans can easily recognize the sense of *administration* in an expression such as *Veteran's Administration*. This observation has led a number of researchers to consider automatic means of extracting multi-word terms from text, terms in which sense variability is more restricted. Unfortunately, the agglutination of adjectives and nouns into longer structures is the general generative method of describing new concepts in English, and it is not clear whether a sequence of words corresponds to an established term or to a description of a transitory real-world situation. In order to extract those phrases which may be considered as terms, document co-occurrence has been used by some researchers (Choueka 1988; Evans *et al.* 1991a; Smadja 1993) as a means of recognizing and extracting prominent noun phrases from a text collection.

Choueka argues that any manually constructed list of two-, three- or four-word terms will not be able to cover the new expressions formed daily in newsprint. He proposes an automatic means of deriving interesting expressions, using the frequency of appearance of the expressions in a large corpus. He proceeds by storing lists of potential expressions appearing more than N times in the corpus. For a ten million word corpus of news wire stories, he used $N = 10$. From these lists he deletes any expression beginning or ending with ''function-words'' or containing frequent words such as numbers, time indicators, or ubiquitous news words like *say* and *reporter*. Then, by automatically overlapping expressions from lists of different lengths, he eliminates some extraneous expressions such as *york times* found as exactly as many times as the longer expression *new york times*. In this way, he finds frequent expressions such as those given in Figure 2.10.

He claims that these simple algorithms find the following types of expressions: personal nouns such as persons, places, organizations, and movie titles; common nouns such as products, organizational titles; and idiomatic and foreign expressions. Although he claims, ''The clear conclusion emerging from even a cursory look at the output is that indeed one can produce surprisingly useful lists," no measure of usefulness is proposed.

```
security council
civil rights
west beirut
federal reserve
executive director
nuclear weapons
federal government
stock exchange
super bowl
foreign policy
san francisco
```

Figure 2.10 Frequently occurring expressions in a large news-wire corpus.

Identification of salient terminology from a given corpus is of great interest (Evans *et al.* 1991a; Choueka 1988) and has been shown to be useful in information retrieval (Hersh *et al.* 1992). But this does not directly address the second half of the word variability problem of relating words not found in the same noun phrases.

2.4.5 Lexical Syntactic Approaches

A few other researchers have started to explore a middle ground between simple word counting and knowledge rich approaches. These researchers accept that a certain level of syntactic analysis is necessary and possible without the need of rich knowledge structures associated with each lexical item. This is our approach throughout the rest of this book. Performing selective natural processing provides the possibility of restricting textual windows to the interior of noun phrases (Ruge 1991), of using contexts such as noun-verb combinations (Hindle 1990), and of recognizing semantic signaling lexical syntactic patterns (Hearst 1992).

CLARIT (Evans *et al.* 1991a) uses the frequency of appearance of syntactically recognized noun phrases in documents as well as in automatically constructed thesauri, in order to identify the salient terms from a document. Candidate terms are generated through exact or partial matches on the thesaurus. Novel terms, whose frequencies in the document warrant notice, are also included among the candidates. Then the terms are ranked on a formula based on *frequency*, *distribution*, and *rarity*. As an evaluation of this technique, CLARIT's performance on a group of ten articles was compared to the terms

container	enclosure, bottle, receptacle, cavity, vessel, tank, pouch
acceleration	deceleration, speed, velocity, inclination, movement, correction
efficient	economical, simple, effective, easy, compact, simultaneous, direct

Figure 2.11 Words sharing the same modifiers in a large number of noun phrases extracted from patent text.

generated by human indexers, and it was found that the terms chosen by CLARIT covered the human indexers' choices better than any individual indexer covered another.

Ruge (1991) implemented a technique which proceeds by first extracting noun phrases from a corpus of 200,000 patent abstracts, and then calculating similarity of heads by comparing the words modifying them. Since each term was sometimes a head and sometimes a modifier, a similarity measure between two terms was developed that took into account the number of shared heads, when the terms were used as modifiers, and the number of shared modifiers when the terms were used as heads. She was able to find relations such as those given in Figure 2.11. Context was restricted to noun phrases only. As an evaluation of the results obtained, Ruge randomly chose 159 words from among the 8257 extracted and had a colleague select synonyms for each. Then a comparison of different similarity measures was performed to see which brought the manually chosen synonyms closest to the top in the automatically generated similarity lists.

Hindle (1990) reports on similar work using noun-verb combinations. He processed 6 million words of 1987 AP news with robust deterministic parsers (Hindle 1989) to extract large numbers of Subject-Verb-Object triples. He then calculated the mutual information between verb-noun pairs. For example, the nouns with the highest associations as objects of the verb *drink* were *bunch-beer, tea, Pepsi, champagne, liquid, beer, wine, water*. As a second order calculation using this mutual information association, he then calculated the similarity between nouns by considering how much mutual information they shared over all of the verbs in the corpus. He was able to produce intuitively pleasing results such as the result that the words most similar to *boat* were *ship, plane, bus, jet, vessel, truck, car, helicopter, ferry, man*. No further evaluation of results was provided. Criticism of using mutual information as a source for detecting similarity surfaced during the 1992 *AAAI Workshop on*

`bruises, ..., broken bones or other injuries`

hyponym(bruise,injury)
hyponym(broken bone,injury).

Figure 2.12 Explicit lexico-syntactic patterns can reveal semantic relations.

Statistically-Based NLP Techniques, when colleagues of Hindle noted that the measure strongly favors rarely appearing words.[2]

Hearst (1992) used lexico-syntactic patterns such as *NP* {, *NP* } *{,} *or other NP* to extract hyponymic relationships between words. Here, *NP* stands for *noun phrase* and the expression *NP* {, *NP* } * stands for one noun phrase followed by any number of noun phrases preceded by commas. If this list of noun phrases is followed by the words *or other* followed by another noun phrase, we can assume that the first set of noun phrases are more specific instances of the final more general term. For example, an expression such as ''bruises, ..., broken bones or other injuries'' which obeys certain other restrictions allows the automatic discovery of the relations given in Figure 2.12. These relations can then be integrated into a hierarchical thesaurus[3], such as has been done for WordNet (Miller *et al.* 1990). As an evaluation of the relations found, the author showed that there was a good overlap between 106 relations that she extracted from *Grolier's American Academic Encyclopedia*, using one such pattern, and a 34,000-word manually-constructed WordNet hierarchy.

In the following chapters, we present our own knowledge-poor system for extracting similar words from corpora of text. We perform a rough syntactic analysis of the corpora in order to extract the contexts by which similarity is judged. Our approach can be seen as a combination of those used by Ruge and Hindle described above, since we use contexts of words both within noun phrases and between nouns and verbs. Similarity is not be be judged by using mutual information but rather by using the Jaccard measure, a similarity metric well known to the information retrieval community. We evaluate our results in a number of ways, including comparison against a manually created resource, as Hearst did to evaluate her work. We show that these techniques address the

[2]The formula for mutual information is $I(x\ y) = log\frac{P(x\ y)}{P(x)\ P(y)}$ where $P(x\ y)$ is the joint probability of the events x and y and $P(x)$ and $P(y)$ are the probabilities of each individual event. The value reaches a maximum when x and y co-occur and are both rare.

[3]We shall examine the conjugation of our method, developed in the next chapter, to Hearst's method for this purpose of thesaurus enrichment in Section 5.2.

two aspects of language variability, as we produce lists of words treating the same concepts, as well as explore the nuances contained in any word via the creation of semantic axes around it. Our work deals principally with individual words, though multi-word terms are discussed at the end. The next chapter presents both the parser that we developed to extract lexical syntactic context for each word and the similarity comparison that we implemented using these contexts.

3

SEXTANT

3.1 PHILOSOPHY

In the last chapter, we provided motivation for the work that we present here by claiming that, in the face of ever greater electronic creation and manipulation of text, the demand for tools to manage and to structure such text will also grow. We have argued that manual approaches to structuring textual knowledge, though useful and promising, cannot keep pace, or be economically justified. In reviewing past attempts at automatic term association, we reviewed work using textual sources whose structure was known, but observed that such work was necessarily limited to a small finite number of sources such as costly man-made dictionaries and thesauri. Such an observation leads us to a discussion of the philosophy of this research, which we outline now.

We are interested in knowing what semantic information can be automatically extracted by computers from unrestricted, large corpora using techniques available today. Unrestricted and large corpora are the corpora of the present and the future. Text is being captured electronically at ever faster rates. The techniques to extract information that we develop here do not rely on hand-built domain knowledge, but use the mass of the text itself as a tool for structuring itself.

In the interest of treating unspecified text, we adopt the philosophy that no domain-specific information should be presupposed or used. In particular, we espouse the following constraints in the interest of producing a domain-independent, robust system that uses today's technology.

- No hand-built domain-dependent knowledge structures
- No interest in extracting information from non-renewable sources such as human-oriented dictionaries, encyclopedias, or thesauri.
- No manually-assigned semantic tags
- No word-specific information in the lexicon (such as linking constraints, valences, transitivity)
- No interest in prescriptive grammars

Though we are voluntarily adopting a knowledge-poor approach, we do not strap ourselves to performing only word counts and string manipulation. Since robust disambiguation and parsing techniques which are independent of the corpus being treated already exist, we can apply these techniques to unrestricted texts. Our system is an attempt to use these techniques on a large scale and to use the regularities of effects apparent in massive data to extract information that otherwise would have to be extracted by hand.

3.2 METHODOLOGY

We present here SEXTANT[1], a complete system that uses fine-grained syntactic contexts to discover similarities between words. The system is based on the hypothesis that words that are used in a similar way throughout a corpus are indeed semantically similar.

Briefly, our system works as follows. After parsing and extraction of syntactic context, we have for each word in the corpus a certain number of objective clues as to its meaning in that corpus. For a noun, for example, we know which nouns and adjectives modify it, as well as verbs of which the noun is the subject or object.

Considering these clues as a word's attributes, similarity measures between words can be calculated. Many similarity measures have been defined and used over the past seventy years (Romesburg 1990). The measures take into account the number of attributes that two objects do or do not share, as well as the importance of these attributes for each word. Words which share a great

[1] The sextant is a navigation instrument which uses the stars as its input. Here SEXTANT stands for Semantic EXtraction from Text via Analyzed Networks of Terms.

number of attributes are found as being similar. Such lists of similar words are one type of output that SEXTANT produces.

What SEXTANT does then is similar to what humans do given an unknown word. The context of the word, i.e., for SEXTANT the other words that have been found to modify it, give a clue to its meaning. Humans make use of much richer contexts, ones involving deeper semantic models of the modifiers and of the discourse structure, whereas SEXTANT currently uses only local lexical clues. SEXTANT collects these clues for each word over an entire corpus and uses them to determine when two words are used in a lexically similar manner throughout the corpus.

In comparison to the more classical techniques, use of syntactic analysis opens up a much wider range of more precise contexts than does simple document co-occurrence, or co-occurrence within a window of words (Phillips 1985). Syntactic analysis allows us to seize more accurately dependencies between words, e.g., to recognize head nouns of phrases, to recognize subjects of verbs, etc., and to develop this information as more precise contexts for word comparison.

Our technique for extracting and using these contexts follows the steps detailed below. Each step is independent of the next, and is of practically linear-time and space complexity, except for the calculation of similarities, which is of quadratic time complexity in the number of unique word-attribute pairs found.

3.2.1 Morphological Analysis, Dictionary Look-up

Raw input text is divided into words by using a regular grammar which we developed (programmed in *lex*) that separates words using spaces and punctuation as delimiters[2]. This grammar embodies certain tokenization rules of English. For example, a certain number of common English contractions such as *'d, 'm, 'll, 're, 've* as well as the genitive *'s* are broken apart without expansion from the word preceding them; a period is considered a separator when it is not in a sequence such as letter - period - letter - period - letter. . . , or in a number.

At this point another simple grammar[3] uses the contextual information of English capitalization to join together sequences of words beginning with

[2] See page 149 in the Appendix 1 for this grammar.
[3] Listed on page 150

an uppercase letter, not appearing after a punctuation mark, as a rapid name recognizer. This is a simple version of a name recognizer which uses the pattern of occurrences of upper and lower case words rather than a list-based system such that proposed by Borkowski (1967) which used lists of proper name markers such *Mr., Mrs., Secretary, Sir, at Large, Acting, ...* as well as lists of common names as a proper name recognizer. After applying this grammar, the original input text has been divided into a number of lexicographical units, called words, one per line.

A morphological normalizer and a 100,000-source-word dictionary, both developed for the CLARIT project (Evans *et al.* 1991b), are used to assign a limited number of syntactic categories, such as `SINGULAR-NOUN`, `PLURAL-NOUN`, `ADJECTIVE`, `GERUND`, `AUXILIARY-HAVE`, `PLURAL-VERB-ACTIVE`, to each word. Words not found in the dictionary are assigned a default category of `UKW`, which is treated in later stages as a noun.

3.2.2 Disambiguation

After dictionary lookup, a word may be labeled with more than one grammatical category. For large natural language processing systems, this is normally the point where a syntactic analysis takes over, producing a parse tree (Sager 1981). In SEXTANT, however, this is not the case. Instead, we use a disambiguator developed by researchers at the Laboratory for Computational Linguistics at Carnegie Mellon University and based on de Marcken (1990) that implements a time linear stochastic grammar based on Brown corpus frequencies. This disambiguator uses two frequencies: (i) the frequency of each word-category pair in the Brown corpus and (ii) the frequency of specific grammatical category sequences in the Brown corpus. The result is the identification of the most probable sequence of grammatical categories (*tags*) throughout a given sentence. Simply assigning its most frequent tag to each word in a corpus results in a 90% tagging accuracy (Brill 1992); this disambiguator that we use has been reported as having a 96% tagging accuracy (de Marcken 1990).

3.2.3 Noun and Verb Phrase Bracketing

Once each word has been disambiguated to a single grammatical category, each sentence is bracketed into noun phrases and verb phrases using another near-linear time algorithm, one originally developed for French texts (Debili 1982;

ADJ :	an adjective modifies a noun	(e.g., civil unrest)
NN :	a noun modifies a noun	(e.g., animal rights)
NNPREP :	a noun that is the object of a proposition modifies a preceding noun	(e.g., measurement along the crest)
SUBJ :	a noun is the subject of a verb	(e.g., the table shook)
DOBJ :	a noun is the direct object of a verb	(e.g., the table was shaken)
IOBJ :	a noun in a prepositional phrase modifying a verb	(e.g., The book was placed on the table)

Table 3.1 Relations extracted by SEXTANT between nouns and other words

Grefenstette 1983) and which we converted to English for SEXTANT. This deterministic bracketer takes information about which tags can start or end a noun phrase, as well as what tags can follow each other within a noun phrase.

In order to perform the bracketing for English text, we have manually constructed a matrix called `CanContinue` with rows and columns representing all of the possible grammatical categories provided by the dictionary. Each cell entry is either a `NO` or a `YES`, indicating whether the sequence of categories represented by a particular row followed by a particular column can be part of a noun phrase. For example, the entry corresponding to `CanContinue(DETERMINER, NOUN)` has a `YES` in it, and the entry corresponding to `CanContinue(DETERMINER,PLURAL-VERB-ACTIVE)` has a `NO`. Two other vectors are used, `CanBegin` and `CanEnd`. Each element of `CanBegin` is `YES` if the corresponding grammatical category can begin a noun phrase, and `NO` if not. `CanEnd` encodes similar information for ending a phrase. For example, `CanBegin(DETERMINER)` is `YES` since it can begin a noun phrase, and `CanEnd(DETERMINER)` is `NO` since a noun phrase cannot end with a determiner. A similar set of matrix and vectors exists for isolating verb phrases. The algorithm for phrase bracketing is found in Figure 3.1.

We have chosen in SEXTANT to extract noun phrases which include prepositions and conjunctions, in order to produce the longest possible complex noun phrases. This was done to extract relations between post-modifying nouns in adjuncts and the preceding noun phrases. Applying the algorithm of Figure 3.1 to a sample text (from a medical corpus) that we use as a running example provides the noun and verb phrases (marked NP and VP in the Figure 3.2).

```
COMMENT Tag[i] contains the grammatical category for word i
currentPhrase = 1
InPhrase = NO
i = 1
while (i <= NumberOfWordsInSentence)
  if (InPhrase == YES) then
     if CanContinue(Tag[i-1],Tag[i]) then
        Phrase[i] = currentPhrase
        next i
     else
        find last word j in current phrase
            for which CanEnd(j) == YES
        currentPhrase++
        i = j+1
        InPhrase = NO
     fi
  else
     if CanBegin(Tag[i]) then
        Phrase[i] = currentPhrase
        InPhrase = YES
     fi
     next i
  fi
endwhile
```

Figure 3.1 Algorithm for Noun or Verb Phrase Bracketing

SAMPLE TEXT:
''It was concluded that the carcinoembryonic antigens represent cellular constituents which are repressed during the course of differentiation of the normal digestive system epithelium and reappear in the corresponding malignant cells by a process of derepressive dedifferentiation.''

NP	it
VP	be conclude
--	that
NP	the carcinoembryonic antigen
VP	represent
NP	cellular constituent
--	which
VP	be repress
NP	during the course of differentiation of the normal digestive system epithelium
--	and
VP	reappear
NP	in the correspond malignant cell by a process of derepressive dedifferentiation
--	.

Figure 3.2 Noun and Verb Phrase Bracketing of Sample Text

```
i = FirstWordInNounPhrase
while (i <= LastWordInNounPhrase)
    if ( Tag[i] in StartSet) then
        j = i+1
        while (j <= LastWordInNounPhrase
               and Tag[j] not Preposition)
           if (Tag[j] in ReceiveSet) then
                 CreateRelation between words i and j
           fi
           next j
        endwhile
    fi
    next i
endwhile
```

Figure 3.3 Left-to-Right Pass over Noun Phrases

```
antigen , carcinoembryonic < ADJ
constituent , cellular < ADJ
digestive , normal < ADJ
epithelium , normal < ADJ
system , digestive < NN
epithelium , digestive < NN
epithelium , system < NN
cell , malignant < ADJ
dedifferentiation , derepressive < NN
```

Figure 3.4 Relations extracted during Left-to-Right Pass

3.2.4 Parsing, Context Extraction

Once the original sentence has been divided into noun phrases and verb phrases, syntactic relations between words within these phrases, and across these phrases, are extracted by a five-pass method that is described in this section. These relations serve as the context of each word in a given corpus of text. All of the relations listed in Figure 3.1 are extracted. These relations hinge upon individual nouns entering into relations with other nouns, with adjectives, or with verbs. Adverbs, numbers, and dates are ignored in this treatment, although such items also could be introduced.

Pass One: Noun Phrases Left-to-Right

In order to extract these relations from bracketed text, first noun phrases are scanned from left to right attaching modifiers such as articles, adjectives and adjectivally used nouns to the farthest noun appearing within the same phrase; the search is interrupted at the end of the noun phrase or at the first preposition.

The algorithm for performing this attachment, given in Figure 3.3, is very simple. Three sets of tags are used: `StartSet` tags which can modify another word, `ReceiveSet` tags which can be modified by a member of the `StartSet`, and a set of tags recognized as `Prepositions`.

A certain number of ambiguities are left unresolved, so this algorithm tends to create more relations than would be produced by a human or by a more

```
i = LastWordInNounPhrase
while (i > FirstWordInNounPhrase)
    if (word i is unattached) then
        if (word i is modified by a preposition)
              j = word before this preposition
        else
              j = i - 1
        fi
        Find the first noun from j downto
                     FirstWordInNounPhrase
    and create a relation between i and this noun
    fi
    i = i - 1
endwhile
```

Figure 3.5 Right-to-Left Pass over Noun Phrases

intelligent system. For example in the phrase *cylinder block manifold* a relation is created between *cylinder* and *block* as well as between *cylinder* and *manifold* and between *block* and *manifold*. Over a large corpus, one of either *cylinder block* or *cylinder manifold* will probably appear more often than the other, and these statistics could be used to decide on the correct relations to retain, although this process has not been implemented in the version of SEXTANT used here[4].

The ADJ and NN relations are recognized during this first pass. Information as to whether a noun is modified by an article or by a preposition is also recorded during this phase. This information is used during the verb attachment phase explained below. Figure 3.4 shows the relations extracted from the given sample text during this pass.

Pass Two: Noun Phrases Right-to-Left

After the first pass, the head noun of any noun phrase or prepositional phrase remains unattached. There may be articles or modifiers or prepositions attached to it, but it remains free to be attached to something else. The purpose of this

[4] See discussion of the problems of noun phrase structure in Section 3.4.3 (p. 65).

his *life* story	------	the story *of his life*
a *dish* cloth	------	a cloth *for dishes*
a *Sussex* man	------	a man *from Sussex*
an *iron* rod	------	a rod *of iron*
life imprisonment	------	imprisonment *for life*
a *Sussex* village	------	a village *in Sussex*
a *gift* tax	------	a tax *on gifts*

Table 3.2 Examples of Premodifying Nouns and their Equivalent Postmodification with Prepositional Phrases.

right-to-left pass is to attach the head nouns of prepositional phrases to a free noun appearing before it. The algorithm for this pass is given in Figure 3.5.

The problem of prepositional phrase attachment can be complex, as can be seen in the well-known sentence: *He saw the girl on the hill with the telescope* in which *with the telescope* may be modifying *hill*, or *girl*, or *saw*. We make no attempt to resolve this problem. In order to do so correctly, semantic information and constraints on the elements being linked are needed. As a heuristic solution, within a complex noun phrase, a prepositional phrase is attached by SEXTANT to the head noun of the preceding phrase[5]. A manually-performed study (Gibson & Pearlmutter 1993) of complex noun phrases extracted from the Brown corpus by SEXTANT reveals that making such a heuristic linking choice provides correct attachment in more than 66% of the cases. In the *telescope* example, SEXTANT retains only the relations between *telescope* and *hill*, between *hill* and *girl*, and between *hill* and *saw*.

While this algorithm proceeds to find a previous noun to which to attach an unattached noun, a record is kept of any prepositions found along the way. If a preposition is discovered before the noun, then a NNPREP relation is created; otherwise a NN relation, for a noun in apposition, is created. The NNPREP relation abstracts away the preposition used. Later, even this NNPREP relation is abstracted away into a general modifying relation. We are concerned with which words modify other words, and not the specific circumstances of this modification. This confusion between pre-modification and post-modification is discussed in Quirk's English grammar: ''In most cases, premodifying nouns correspond to postmodification with prepositional phrases'' (Quirk *et al.* 1985, p. 1330). Examples from this grammar are given in Figure 3.2.

[5] The prepositional phrase is also sometimes attached to a preceding verb, see page 44.

```
course , differentiation < NNPREP
differentiation , epithelium < NNPREP
cell , process < NNPREP
process , dedifferentiation < NNPREP
```

Figure 3.6 Relations extracted during Right-to-Left Pass

```
repress , antigen < DOBJ
antigen , represent < SUBJ
represent , constituent < DOBJ
reappear , cell < IOBJ
cell , correspond < SUBJ
```

Figure 3.7 Relations extracted during Verb Passes

During this pass over the sample text, the relations given in Figure 3.6 are extracted.

Pass Three: Verb Phrases Right-to-Left

After the first two passes, there are usually unattached nouns before and after each verb phrase. The next two passes attempt to attach verbs to these nouns as their subjects and objects.

Before Pass Three, each verb phrase is analyzed to find the head verb and to determine if the phrase is active or passive. This analysis is simple: Trace the verb phrase to its last verb, this becomes the head verb. A verb phrase begins as ACTIVE. If an auxiliary verb form of *be* is found the verb phrase is PASSIVE. If a progressive verb (other than *being*) is found, then the phrase becomes ACTIVE. If the head verb is a form of *to be*, then the verb phrase becomes ATTRIBUTIVE.

Once the verb phrase has been analyzed, the SEXTANT parser searches for the first free noun (not attached to another noun) before the verb phrase which becomes the subject (or direct object if the verb phrase is PASSIVE).

Pass Four: Verb Phrases Left-to-Right

During Pass Four, a similar search takes place to find an unattached noun which becomes the direct object of an ACTIVE verb phrase. The first head noun of a prepositional phrase after the verb phrase becomes the IOBJ of the verb. The IOBJ is not always the indirect object as linguists use the term; it can also be a noun that modifies the verb in some general sense. For example in the phrase ''give to the doctor'', *doctor* is the indirect object of the bi-transitive verb *give.* In the phrase ''reappearing in the cell'', *in the cell* modifies the verb in a locative sense. The relation IOBJ extracted by SEXTANT confuses these two senses. IOBJ can be interpreted as a general tertiary relation between a verb and a noun.

Pass Five: Progressive Participles

A fifth pass goes through the text trying to attach progressive verbs, ending in *-ing*, to potential subjects and objects. Progressives can appear in noun phrases as nouns ''the heating of the solution'', as adjectives ''the heating pad'', or between noun phrases ''the element heating the water''. Correct treatment of progressives is a corpus-dependent problem and is left unattacked by SEXTANT. Instead a simple heuristic is implemented. We consider a progressive following a determiner or a quantifier to be a noun; all others are considered as progressive verbs whose subjects precede and objects follow.

This fifth pass follows the same algorithms as Passes Three and Four, but with a relaxation of the constraint that the subjects and objects found be unattached. During the last three passes over the previously given text, the relations given in Figure 3.7 are extracted.

The original sample with the relations[6] extracted by SEXTANT is given in Figure 3.8. Note that the information about `ADJ`, `NN`, or `NPREP` relations are stripped away, cf. page 42, and only the information concerning verbs is retained.

[6] These relations will be used here to compare words to each other. Grishman & Sterling (1992) used such relations to improve parsing.

SAMPLE TEXT:

''It was concluded that the carcinoembryonic antigens represent cellular constituents which are repressed during the course of differentiation of the normal digestive system epithelium and reappear in the corresponding malignant cells by a process of derepressive dedifferentiation.''

```
antigen carcinoembryonic
antigen repress-DOBJ
antigen represent-SUBJ
constituent cellular
constituent represent-DOBJ
course repress-IOBJ
course differentiation
digestive normal
epithelium normal
system digestive
epithelium digestive
epithelium system
differentiation epithelium
cell correspond-SUBJ
cell malignant
cell reappear-IOBJ
cell process
dedifferentiation derepressive
process dedifferentiation
```

Figure 3.8 Simplified Contexts for Each Word

3.2.5 Discussion about Parser Results

As can be expected from the simplicity of the algorithms for disambiguating the text and for extracting and analyzing phrases, errors may appear among the otherwise acceptable list of relations extracted. For example, the true subject of *reappear* should be *constituents*, which is absent from Figure 3.8. Figure 3.7 shows *antigen* as the direct object of *repress* whereas the true syntactic object should be *constituents*.[7] We tabulated the performance of the parser over a sample of sixty sentences and found that of the 440 relations extracted, 75% were correct. The errors occur for a number of reasons, e.g., words missing from the lexicon and mislabeled as a noun, words incorrectly tagged by the disambiguator, and parser limitations, most notably due to the presence of conjunctions. The 75% rating can be compared to much more elaborate parsers (Grover *et al.* 1993; Usioda *et al.* 1993) which can achieve correct attachment rates of 80-85% on correctly tagged text.

Our simple algorithms are used for reaping information from great quantities of text, and not for providing a model of human competence. They have the advantage of being very fast, since no backtracking, recursion, or maintenance of possible parses is involved. Although a great many serious linguistic problems are not addressed, such as anaphora resolution, multi-word verbs, garden paths, etc. (Smith 1991), SEXTANT's parser does provide correct results for the simpler and more common constructions[8]. In this sense, it can be seen as an improvement over coarse-grained windowing techniques which connect any word to any other appearing within a textual windowing. When SEXTANT makes an error of attachment, it defaults in some sense to this coarse technique since it then attaches together two non-syntactically related words appearing within a small window just as a windowing technique would. As we shall see in Section 3.2.7, even with these imperfections, over a large enough corpus, useful domain knowledge can be generated by SEXTANT.

3.2.6 Similarity Calculations

Once the syntactic analysis of the corpus is performed, each word in the corpus possesses a certain quantity of context which SEXTANT uses to judge word similarity.

[7] The original phrase ''It was concluded that the carcinoembryonic antigens represent cellular constituents which are repressed during the course of differentiation ...'' seems however to indicate that the constituents and the antigens are co-referent.

[8] Such constructions correspond to what can be obtained by a *syntactic sketch*, one of the lowest levels in a more complex dynamic relaxation parser (Chanod *et al.* 1993).

```
for each noun i in the corpus
    for each noun j <> i in the corpus
       use a similarity measure to calculate
             the distance between noun i and noun j
    sort the nouns according to the similarity to noun i
    retain the closest N words to noun i
endfor
```

Figure 3.9 Word-by-word similarity comparison

Similarity between objects based upon shared attributes has been widely studied in many experimental and social sciences. SEXTANT implements a wide variety of similarity measures described in (Romesburg 1990), a clear introduction to similarity calculation. Similarity comparison over the word contexts extracted in Section 3.2.4 is performed by the algorithm given in Figure 3.9.

For the moment, let us consider nouns, although the same techniques is applied later to modifiers or verbs. Each noun found in the text is considered an object, and the words that are found to modify it are considered its attributes. A noun can be modified by an adjective (`ADJ`), by another noun (`NN` and `NNPREP`), or by a verb (`SUBJ`, `DOBJ`, and `IOBJ`), and these modifications are taken to be the known attributes of the noun.

Jaccard Similarity Measure

The similarity measure that seems to produce the best results in SEXTANT is a weighted Jaccard similarity measure, also known as the Tanimoto (1958) measure. The binary Jaccard measure between two objects m and n is the number of shared attributes divided by the number of attributes in the unique union of the set of attributes for each object:

$$\frac{Count(\{Attributes\ shared\ by\ object_m\ and\ object_n\})}{Count(\{Unique\ attributes\ possessed\ by\ object_m\ or\ object_n\})} \tag{3.1}$$

Let's give an example. Suppose that we are comparing two nouns *dog* and *cat* possessing the explicit textual attributes derived as described in Section 3.2.4

```
dog pet-DOBJ
dog eat-SUBJ
dog shaggy
dog brown
dog leash
cat pet-DOBJ
cat pet-DOBJ
cat hairy
cat leash
```

$$\frac{Count(\{leash, pet_{DOBJ}\})}{Count(\{brown, eat_{SUBJ}, hairy, leash, pet_{DOBJ}, shaggy\})} = \frac{2}{6} = 0.333$$

Figure 3.10 Comparing 'dog' and 'cat' via textually derived attributes and a binary Jaccard measure of similarity

and shown in Figure 3.10. Suppose that the word *dog* has 5 attributes and *cat* has 3 attributes, one of which appears twice. In a strictly binary Jaccard measure, the similarity of *cat* and *dog* would be 0.333.

Weighted Jaccard Measure

Moving from a binary to a weighted measure can be done in many ways. We have found it useful to weight attributes using a log-entropy weighting that has been shown to improve document retrieval, a related problem, in Dumais (1991). Each attribute is assigned a global weight between 0 and 1 depending upon how many different objects with which it associates, and how often it appears, using the Formula 3.2. A higher global weighting means that the word appears less often in the corpus. Our formula for the weighted Jaccard similarity measure between two objects $object_m$ and $object_n$ is given in Formula 3.5.

$$1 - \sum_j \frac{p_{ij} \log(p_{ij})}{\log(nrels)} \tag{3.2}$$

$$p_{ij} = \frac{absolute\ freq\ of\ attribute_j\ with\ object_i}{total\ number\ of\ attributes\ for\ object_i} \tag{3.3}$$

brown	0.9	eat-SUBJ	0.7
hairy	0.85	leash	0.75
pet-DOBJ	0.6	shaggy	0.8

$$\frac{0+0+0+0.75+0.6+0}{0.9+0.7+0.85+0.75+0.79+0.8} = 0.28$$

Figure 3.11 Example weighted attributes and their use in the weighted Jaccard similarity measure between 'dog' and 'cat'

$$nrels = the\ total\ number\ of\ relations\ extracted\ from\ the\ corpus \tag{3.4}$$

$$\frac{\sum_{unique\ attributes} \min(weight(object_m, attribute), weight(object_n, attribute)}{\sum_{unique\ attributes} \max(weight(object_m, attribute), weight(object_n, attribute)} \tag{3.5}$$

Note that when the weights are restricted to 0 and 1 this last formula is equivalent to the previous binary Jaccard formula, although this is by no means the only way in which to generalize the binary formula to the weighted case.

In order to show how the weighted Jaccard similarity is calculated in SEXTANT, let us suppose in this *dog - cat* example that the global weights of the attributes, when calculated over a whole corpus, are those given in Figure 3.11.

Before calculating similarity, a local weighting is also given to each object-attribute pair. If an attribute appears more than once for word, as *pet-DOBJ* does for *cat*, then the weight of that attribute, in this log-entropy scheme, is multiplied by the log of its frequency for that word. For example, since *cat* has the attribute *pet-DOBJ* twice, the weight of that attribute for *cat* is its global weight multiplied by its local weight, which gives $0.6 \log(2+1) = 0.79$. The value of the attribute *pet-DOBJ* for *dog* would be $0.6 \log(1+1) = 0.6$. Now, although *cat* and *dog* share the attribute *pet-DOBJ* equally in the binary

case, in the weighted case the weight of this attribute is greater for *cat*. This logarithm is applied to temper the effect of frequently appearing modifiers in the Jaccard calculation.

For our *cat* and *dog* example, this weighted similarity measure gives 0.28, as seen in Figure 3.11. In other words, they are a little less similar than in the binary case, since *cat* possesses one attribute to a different degree than *dog*. As can be seen in this example, when the number of attributes shared is small, the similarity measure is small. The drop off from the perfect match yielding a 1.0 is rapid. What is important is the relative value of similarity of objects, and the ranking of objects using these relative values is what we exploit in SEXTANT.

3.2.7 Results

Once the similarity of each word is calculated and the results sorted, SEXTANT produces a list of the most similar words for each word in the corpus. In the Appendix 3 and for each corpus described in Appendix 6, we present the results of applying SEXTANT to large corpora. These results are presented in the following format:

<word> [<number of attributes>] <list of most similar words >.

The <list of most similar words>[9] is divided into groups of words whose similarities to the leftmost <word> are within 0.01 of each other. For example, Figure 3.12 presents a sample of the lists extracted by SEXTANT from the `MED` corpus, the word *tissue* was closest to *cell*, with a similarity measure of 0.06, while a number of other words *growth, cancer, liver,* and *tumor* were found to be less similar, all with a measure of around 0.04. The words *resistance, disease, lens, serum,* and *lesion* were less similar at 0.03.

As a detailed example, we study the list extracted for the word *cause* in the `MED` corpus. The `MED` corpus, from which the sample text cited in previous sections was drawn, is a 1 megabyte corpus often used as a testbed in Information Retrieval. The word *cause* appears 151 times in this corpus. Eighty-three times it is recognized as a noun. In these cases, it was found to be modified by sixty-seven unique words, some of which are shown in Figure 3.13.

[9] These output lists have arbitrarily been truncated at ten words throughout the SEXTANT system.

word [*Contexts*]	*Groups of most similar words*
tissue [350]	cell | growth cancer liver tumor | resistance disease lens serum
treatment [341]	therapy | patient administration case response | result effect
concentration [339]	level content | excretion value | rate ratio metabolism synthesis
defect [338]	disturbance case malformation | regurgitation type response
rat [331]	animal mouse | dog mice level | infant kidney day rabbit group
method [298]	technique | procedure test | mean result study | group treatment
pressure [286]	flow | volume | artery obstruction rate tension | serum sinus level
growth [284]	tumor | tissue increase effect development protein response
test [284]	technique method reaction | response study therapy observation
tumor [260]	carcinoma growth | cancer lesion | sarcoma tissue effect lung

Figure 3.12 Sample of Similarity Lists Extracted by SEXTANT from MED.

cause arachnoiditis
cause basic
cause cell
cause child
cause clarify-DOBJ
cause clear-DOBJ
cause common
cause concern
cause concern-IOBJ
cause constriction
cause death
cause deficiency
cause dehiscence
cause dehydrogenase
cause determine-DOBJ
cause different
cause discuss-DOBJ
cause discuss-SUBJ
cause establish-DOBJ
. . .

. . .
cause propose-DOBJ
cause rare
cause recognize-DOBJ
cause red
cause regurgitation
cause reveal-DOBJ
cause selenocystine
cause series
cause single
cause stability
cause suppose-DOBJ
cause suspect-IOBJ
cause symptom
cause thymectomy
cause uefaincrease
cause ulceration
cause uncertain
cause unknown

Figure 3.13 Sample of words modifying *cause* in the MED corpus.

attribute	*Global Weight*
basic	0.37186
determine-DOBJ	0.23723
hepatitis	0.30438
hydrocephalus	0.26803
jaundice	0.31582
possible	0.26803
uncertain	0.48928
unknown	0.39321

Figure 3.14 Corpus derived weights of attributes shared by *cause* and *etiology* in the MED corpus.

These words form the attributes of the word *cause* that are used to find similar words in the corpus. When the whole corpus is analyzed by the methods described in the two preceding sections, 4954 different nouns are found to possess 9209 different attributes. When the similarity measure described in the previous section was applied to all of the extracted nouns, *cause* is found to be closest to the word *etiology*, which itself possessed thirty-five unique attributes. Together, the two words share the attributes given in Figure 3.14 with their corpus-derived weights (cf. page 46).

The Jaccard calculation of shared attributes between *cause* and *etiology* yields a similarity of about 0.07, which is small as an absolute measure, but larger than the similarity of *cause* to any of the 4953 candidates in this corpus. The next closest word using this similarity measure is *explanation*, which shares the attributes {*possible discuss-DOBJ phenomenon likely find* } and is measured at a similarity of about 0.05. A few candidates come in at a similarity measure of about 0.4: *incidence, nature, component, evidence, dehydrogenase, feature, diagnosis, pattern*. In this corpus, more than 2400 words share at least one attribute with *cause*. Hence, using SEXTANT's weighted Jaccard measure over these words produces non-null similarities, however, very few words share many attributes with *cause*. Although *etiology* is calculated as the most similar word to *cause*, the words *evidence* and *diagnosis* actually share more attributes with *cause* than *etiology* does but they are also words which appear with much greater frequencies in the corpus than *etiology*. Thus, the number of the attributes that are not shared, a value appearing in the denominator of the Jaccard measure, is also greater for *evidence* and *diagnosis*, reducing the similarity measure.

no cause could be determined in thirteen cases, which were, therefore, labeled as 'primary' amyloidosis .	after an acute encephalopathy, the etiology of which could not be determined
. . . that one of the possible causes of so-called giant cell hepatitis is an inborn error of metabolism.	the etiology of giant cell hepatitis is still unknown.
the possible causes for this phenomenon are being discussed .	the sex incidence and association of other congenital anomalies is discussed in relation to a possible etiology.
in 30% of 301 patients with cirrhosis, the cause was uncertain.	the etiology of autism is uncertain.
the basic cause being intracranial hemorrhage in the perinatal period. the most common cause of hydrocephalus is arachnoiditis followed by congenital anomalies.	intellectual capacity could not be correlated with the basic etiology of the hydrocephalus. . . . apparently irrespective of basic etiology of the hydrocephalus.
red cell glucose-6-phosphate dehydrogenase deficiency--a newly recognized cause of neonatal jaundice and kernicterus in canada	it is interesting to re-examine current concepts of the etiology of physiologic jaundice with this diagram in mind.
. . . four possibly due to traumatic birth or neonatal asphyxia, and four from an unknown cause	these distinctions and classifications are thought to be controversial, so long as the etiology of early infantile autism is unknown.

Figure 3.15 Corpus evidence used by SEXTANT for calculating similarity of *cause* and *etiology* in MED.

Comparing complete sets of attributes (which include non-shared attributes for two words, with respect to each other) does reveal words which seem semantically similar. However, no single shared attribute is sufficient to determine that two words are similar. Note the evidence presented in Figure 3.15 for considering that *cause* is similar to *etiology* in this corpus[10].

This effect of many words interacting to fix a meaning is reminiscent of the ''lexical field'' concept developed by Jost Trier in the early part of the twentieth century. Lexical fields (Ullmann 1962; Lehrer 1974) were blocks of human experience which were sliced up by a number of lexical units which covered that field. For example, the color names covered the lexical field of color, and each name derived its meaning from the whole system and its position in that system. This idea that each word influences the meaning of the words around it finds its echo in what SEXTANT extracts. If each single word, when it modifies another word, somehow delimits the meaning of that word, it becomes likely that words which are modifiable by the same groups of words are delimited in the same way, and are therefore similar. Lexical field theory did not find favor with the present generation of generative linguists; still we believe that it provides an explanatory image of what happens in SEXTANT.

3.3 OTHER EXAMPLES

Our semantic extraction technique was applied to a wide variety of corpora, each described separately in the Appendix. Here we give some examples of how SEXTANT is able to extract corpus-specific relations. For example, the corpus `ANIMALS` consists of 756 animal articles extracted by us from *Grolier's Encyclopedia*. This extraction was made by using a man-made thesaurus which listed a great number of animals and by matching these names to encyclopedia article titles. A manual verification eliminated non-animal articles. Samples of this text are presented in the Appendix. Below is the beginning of the article dealing with zebras:

> The zebra is a distinctly striped, hoofed mammal in the genus Equus (which includes the horse and the ass) of the family Equidae. The head and body are about 2.3 m (7.5 ft) long, with a 56-cm (22-in) tail ending in a tuft of hair. The height at the shoulder is 1.2 to 1.5 m (4 to 5 ft); the weight reaches up to 346.5 kg (770 lb). The smooth whitish

[10] This is the `MED` corpus described in the Appendix. The original text of this corpus was only available in lower case.

word[*contexts*]	*Groups of closest words*
zebra [12]	rhinoceros | hyrax | bear shrew seal | order genera mammal adult duck
mammal [157]	rodent animal lizard | snake form crab life reptile bird deer
coat [245]	fur | plumage hair | color feather patch | stripe marking tail yellow
genus [35]	tribe subclass relative corvidae snipe | vulture passeriforme sparrow rodent
horse [134]	duck pony terrier sheep | number dog trout pigeon animal rat
ass [8]	bear deer duck monkey horse | crab worm year snake female
family [401]	species | genera subfamily range form | order variety size number nest
head [265]	bill tail body | wing back coat color fin ear eye
tail [385]	bill wing | leg head | body coat ear foot | toe fur
tuft [15]	underside | shape | patch marking feather | color side wing structure part
hair [109]	coat fur | marking mane stripe feather band | pigment plumage spot
height [37]	length weight | speed | centimeter north size | surface tail range area
weight [104]	length | speed height | maturity total size shoulder | litter density surface
band [71]	stripe | marking spot | patch plumage | shade bar green brown fur
stripe [87]	band | marking spot | patch | bar plumage underside | brown tan side
pattern [85]	feature adaptation fur characteristic plumage green | social wing change
species [1404]	bird fish | family group form | animal insect range snake | male

Figure 3.16 Similarity lists extracted from ANIMALS for the nouns in the definition of ‘‘zebra’’.

> or tawny coat is marked with striking dark brown or black bands or stripes, which have a distinctive pattern in each of the three species of zebras. . .

When SEXTANT is passed over this entire corpus, the nouns in the above sample of text yield the similarity lists given in Figure 3.16. As can be seen, animals are classed mostly with other animals, body parts are classed with other body parts, features are classed with other features, measures are classed with other measures, and generic words are classed with other generic terms.

In Figure 3.16, it can be seen that words possessing more context produce lists whose similarity is more apparent and in which we can have more confidence. It can also be seen that, although the relation between a word and the closest few most similar words can be divined, ‘‘ringers’’ appear among less similar words. For example, for the word *horse*, *number* is produced as the fifth closest word[11].

[11] According to SEXTANT, ‘horse’ and ‘number’ share the following attributes: *small animal use-DOBJ gray extinct breed-IOBJ name great time increase-SUBJ breed-DOBJ farm today eater draft decrease-SUBJ*. The presence of ‘number’ as similar to ‘horse’ can be considered as noise generated by the summary syntactic analysis performed by SEXTANT. In fact, the word ‘number’ is really a pseudo-quantifier in expressions such as *a number of* . . . whereas the syntactic analysis in SEXTANT considers ‘number’ as it would any other noun and generates

word[*Contexts*]	*Groups of closest words*
average [102]	hit total | situation rest avg gene level difference hitter oba
baseball [140]	sport | time mistake hitter series chance defense atlanta day canada
catcher [44]	fielder | race arm | hitter stuff shortstop work outfielder bell park
contract [46]	market offer projection winter roster | book puckett failure assignment
field [78]	fielder place corner | variability possibility range town glove bonehead
hit [214]	run play average | hr inning hitter pitcher time shot guy
hitter [176]	pitcher | chance fielder bullpen play manager player | pitch baseball
homer [46]	fly hr shot | stanton effect table slam pinch part dugout
jay [205]	atlanta twin pirate | lot toronto time play blue-jay pitcher people
loss [42]	victory | whera couple work anthem win woof red-sox hand | pirate
money [34]	offer | club buck fuss deal job fine | surprise franchise van-slyke
mvp [43]	cy-young leader | notch west respect history ring proximity race cf
nlc [27]	ws division | alc buchelle gold-glove serie | order total stanley-cup
park [35]	dugout leader history take | major office starter shortstop comebacker
performance [73]	record | offense stat base ability hr win clutch talent point
pitcher [280]	hitter | player season time win | lot number run series ball
plate [65]	hr dugout | ground shot mound park rightie room turf homer
player [429]	pitcher | play hitter fan number point manager time run | season
staff [55]	bullpen toronto start sox | blunder job matchup win coach good
stat [99]	number point performance record | average void guy change week
system [48]	owner manager cito | matter technique management club office
tv [23]	picture video screen correlation | affiliate live network nlc
twin [95]	jay blue-jay brave | series way nlc atlanta guy toronto pittsburgh
world-series [82]	ws pennant deal series nlc fun alc division | red-sox time
ws [34]	nlc | world-series alc gold-glove serie championship | posting trophy

Figure 3.17 Baseball related terms in the BASEBALL corpus and their associated similarity lists.

As another example of applying SEXTANT's similarity calculation to a specific corpus, we downloaded a body of extremely disparate and disconnected text, including things such as tables and C programs from a newsgroup (rec.baseball). The only cleaning done on the text was removal of address headers and signatures. Samples of this corpus can be seen in the Appendix. In Figure 3.17 we present the similarity of baseball-specific words present in this corpus.

Although there is quite an amount of noise in the above relations, due in part to the loose structure of the original newsgroup text, the words *players, catchers, fielders, pitchers, hitters* are associated with each other, as are end-of-year competitions such as *nlc, ws, world-series, alc,* and baseball teams such as *twins, jays, braves, blue-jays, atlanta, pirates*.[12] Also, parts of the playing field, such as *plate, dugout, park, field, corner* are brought together, as are things relating to scoring such as *stat, point, performance, record, hit, run*.

3.4 DISCUSSION

3.4.1 Time and Space constraints

One might be concerned about the tremendous space involved in storing not only the words appearing in a large corpus, but also the relations between them. Although this space is theoretically the square of the number of words in the corpus, in reality the number of relations extracted is much smaller. In Figure 3.18, we give the space required for some of the corpora to which we have applied SEXTANT. The table gives the size of each corpus, the number of individual words, the number of unique words, the number of word-attribute pairs, the number of characters in the word-attribute pairs, and the number of unique word-attribute pairs. It is this last number which is the determining factor for the speed of SEXTANT's similarity calculations.

It can be seen that the space needed to store the data for the calculation of the similarity relations is of the same order as that needed to store the original text. As for time, the calculation of similarity of words, given their attributes,

---

all the attendant attributes. If a better syntactic analysis module were plugged into SEXTANT such noise could be reduced.

[12] Note that terms such as 'world-series' and 'blue-jays' do not appear hyphenated in the original text and also do not appear in the lexicon. These terms are recognized as units by the proper name recognizer described in Section 3.2.1 and in Appendix 1.

corpus	*text size*	*words*	*unique*	*objects*	*attribs*	*pairs*	*data size*
ADI	38 K	5,500	1,500	537	931	2030	44 K
AI	2800 K	387,000	25,000	6549	11638	93739	3116 K
AIDS	2800 K	458,000	22,000	8041	14216	100593	3273 K
ANIMALS	1200 K	200,000	18,000	7007	10027	48085	1087 K
BBALL	950 K	190,000	16,000	5230	7260	28338	586 K
BROWN	6200 K	1,000,000	46,000	22079	32360	235051	1367 K
CACM	1300 K	193,000	9,700	4760	8070	46027	1283 K
CISI	1300 K	204,000	12,000	4654	8402	48475	1288 K
CRAN	1600 K	260,000	11,800	3057	6594	49489	1925 K
HARVARD	3900 K	665,000	50,600	17659	23587	136251	3488 K
JFK	1820 K	360,000	18,600	7249	11446	61081	1522 K
MED	1000 K	187,000	14,500	4966	9221	47670	1229 K
MERGERS	5200 K	458,000	45,500	18739	23934	183446	5718 K
MOBY	1000 K	244,000	19,600	7261	11860	47492	888 K
NEJM	1000 K	184,000	8,700	2699	5228	28578	1075 K
NPL	3200 K	490,000	23,600	5101	9646	115108	254 K
SPORTS	6000 K	1,100,000	88,000	25877	33539	229562	5450 K
TIME	1500 K	287,000	22,000	9600	16360	80190	1643 K

Figure 3.18 Space Constraints for Eighteen of the Corpora Treated.

involves comparing the attributes of each word to those of every other word. If we have N words and M possible attributes, we would theoretically have $O(MN^2)$ comparisons to make, since for each pair of words we would make up to M comparisons. But in fact the matrix of words-by-attributes is very spare so the actual complexity is much smaller. We store the matrix as a linked list, and use bit signatures (Faloutsos 1992) as a quick check of whether words have any attributes in common. Figure 3.19 gives an idea of the speed of the algorithm. For each corpus, we give the number of real time minutes that it took to perform the steps on a DEC-5820[13].

The column labeled *Label* shows the time needed to convert the raw text into words and perform the dictionary lookup of each word. This time includes running the text through an *awk* program to produce tokens and recognize some proper names, loading the lexicon, running a morphological analyzer developed for CLARIT (Evans *et al.* 1991b), and producing an intermediate file of tagged text. The column *Disamb* shows the time used to disambiguate

[13] This DEC-5820 was graciously lent to us by the Laboratory of Computational Linguistics, directed by Dr. David Evans, at Carnegie Mellon University. The real-world time measures given here were on a machine that is the file server for the Laboratory. Since the measurements were made during a normal academic week some jobs running during the day may have been slower than those running at night.

	S P A C E				*T I M E (in real world minutes)*			
corpus	*text size*	*objects*	*attribs*	*pairs*	*Label*	*Disamb*	*Extract*	*Sim*
ADI	38 K	537	931	2030	1.11	.48	.11	.16
AI	2800 K	6549	11638	93739	12.63	15.50	4.02	68.31
AIDS	2800 K	8041	14216	100593	6.68	13.53	4.28	105.65
ANIML	1200 K	7007	10027	48085	6.58	11.33	3.80	52.41
BBALL	950 K	5230	7260	28338	4.85	5.21	1.21	10.40
CACM	1300 K	4760	8070	46027	3.51	6.81	1.70	15.65
CISI	1300 K	4654	8402	48475	3.41	5.61	1.76	18.23
CRAN	1600 K	3057	6594	49489	1.76	9.88	2.61	17.61
HARV	3900 K	17659	23587	136251	19.08	22.81	10.30	219.36
JFK	1820 K	7249	11446	61081	2.12	10.13	2.43	61.07
MED	1000 K	4966	9221	47670	1.61	4.91	2.50	30.16
MERG	5200 K	18739	23934	183446	13.93	25.30	8.16	336.13
MOBY	1000 K	7261	11860	47492	3.93	7.00	2.01	29.50
NEJM	1000 K	2699	5228	28578	3.05	5.08	1.61	7.93
NPL	3200 K	5101	9646	115108	3.31	5.43	5.13	75.90
SPORT	6000 K	25877	33539	229562	17.21	31.73	9.25	542.13
TIME	1500 K	9600	16360	80190	1.98	14.30	5.45	84.50

Figure 3.19 Time Constraints for Treating Corpora with SEXTANT.

this tagged text using the algorithm[14] described in (de Marcken 1990). The column *Extract* shows the time needed by the programs we developed to bracket phrases, parse them and extract the contexts for each word in the text, as explained in Sections 3.2.3 to 3.2.4. The column marked *Sim* shows the time needed to compare the contexts of each word in the text. Words appearing only one or two times in the corpus were not included in the comparisons, although they were counted in the column marked *objects*. This reduced the number of objects by about half, but the number of pairs by about 10%.

As can be seen in Figure 3.19, the costliest part of SEXTANT's processing is the calculation of similarity, which is $O(pairs^2)$, since similarity comparison is essentially comparing lines of a matrix[15].

[14] This algorithm was implemented in C by David Leberknight of the Laboratory of Computational Linguistics, Carnegie Mellon University.

[15] The exact relation between the number of *pairs* and the running time in minutes follows the plot of the equation: $minutes = (0.000097 pairs + 0.22)^2$ We found this equation by plotting the number of pairs against the square root of the number of minutes of execution time, then doing a linear regression.

3.4.2 Stability of Results

It is important to judge the stability of the methods employed here, to know whether the similarities discovered by this method change easily when new text is added, or whether the results become more and more certain as more text is treated. It is clear that if text from a completely different domain is mixed into a corpus that contexts of certain words will change. Imagine for example a corpus where *caterpillar* is a machine mixed in with text in which *caterpillar* is an insect. The context of that word would be altered by the new usages and its similarity to other words would change accordingly. On a smaller scale, however, each use of a word even within one domain alters in some sense its meaning. Word meanings have a tendency to drift over time, just as pronunciation does, through infinitesimal changes in usage. We can hardly measure historical changes with the corpora and the techniques that we possess, but we can measure the stability of the results we obtain within one corpus by measuring the changes in the similarity lists produced by running SEXTANT over different percentages of the corpus.

In order to measure stability, we ran the similarity extraction modules of SEXTANT over 50%, 75%, 85%, 95%, and 100% of a 2.8 megabyte corpus of abstracts about AIDS. At each change in percentage we extracted those words for which new context was added and compared the similarity lists before and after the additions. We measured when there was substitution in the most similar word, among the first two most similar words, among the first five most similar words, and among the first ten most similar words. The results of these changes are presented in the following two graphs.

Figure 3.20 shows the results of adding 1 to 20 new contexts to words from this corpus. The words are divided into frequency groups. For example, there are 1792 words possessing fewer than 20 contexts before the new contexts are added; there are 1417 words possessing 20 to 49 contexts; 730 possessing 50 to 99; 458 words possessing 100 to 199 contexts; 221 possessing 200 to 499 contexts; and 14 possessing from 500 to 6678 contexts. To each word in each group were added from 1 to 20 new contexts, corresponding to more text treated. The words considered to be similar by SEXTANT before and after the addition of this text were compared. The four bars above each frequency group show the results of this comparison.

The bar above $1'$ shows the percentage of words in that frequency group for which the most similar word did not change after adding more context. For example, for words possessing 20 to 49 contexts, the effect of adding text

resulting in 1 to 20 more contexts to these words was that 80% of the most similar words remained unchanged[16].

The bar marked $2'$ shows the percentage change in the contents of the two closest words. If the first two words merely switched places after new context is added, the switch is considered as no change. For example, for the 221 words possessing between 200 and 499 contexts, adding 1 to 20 new contexts results in no change in the two closest words, although for two of the words in this frequency group, adding this new context reversed the order of the two closest words to them, reflected by the 99% over the $1'$ bar.

The bars marked $5'$ and $10'$ reflect the percentage changes in the five closest words and in the ten closest words, for the words in each frequency group. For example, the bar marked $10'$ above the frequency group marked 50-99 corresponds to 76%. This means that if you compare the ten closest words to each of 730 words in this frequency group *before* new text is added to the ten closest words to each of 730 words *after* the new text is added, then 76% of these words are the same.

Figure 3.21 shows the results of adding 20 to 50 new contexts and Figure 3.22 shows that of adding 50 to 100 new contexts. In all three graphs it can be seen that more frequent words are more stable; as the number of contexts rises, the change in closest words becomes rarer as more context is added. Across the graphs there is a similar phenomenon. Consider the frequency groups in each graph for which the context is doubled. In Figure 3.20, 56% of the frequency group 1-19 retain the most similar word when their context is doubled; in Figure 3.21, 64% of the group 20-49 retain the most similar word when their context is doubled; in Figure 3.22 85% of the words in the group 50-99 retain their closest word as their context is doubled. The tendency seems to show that once the context becomes well established, small changes rarely modify the results. Moreover, as the number of contexts recognized grows, changes become rarer and rarer even when the context doubles, suggesting a growing stability, as is visually apparent in the graphs.

[16] Since there were 1417 words in this frequency category, this means that 80% or 1134 words retained the same word as being most similar after their contexts were increased by about 50%. On the other hand, for 283 words, adding this much new context produced a 'new' most similar word.

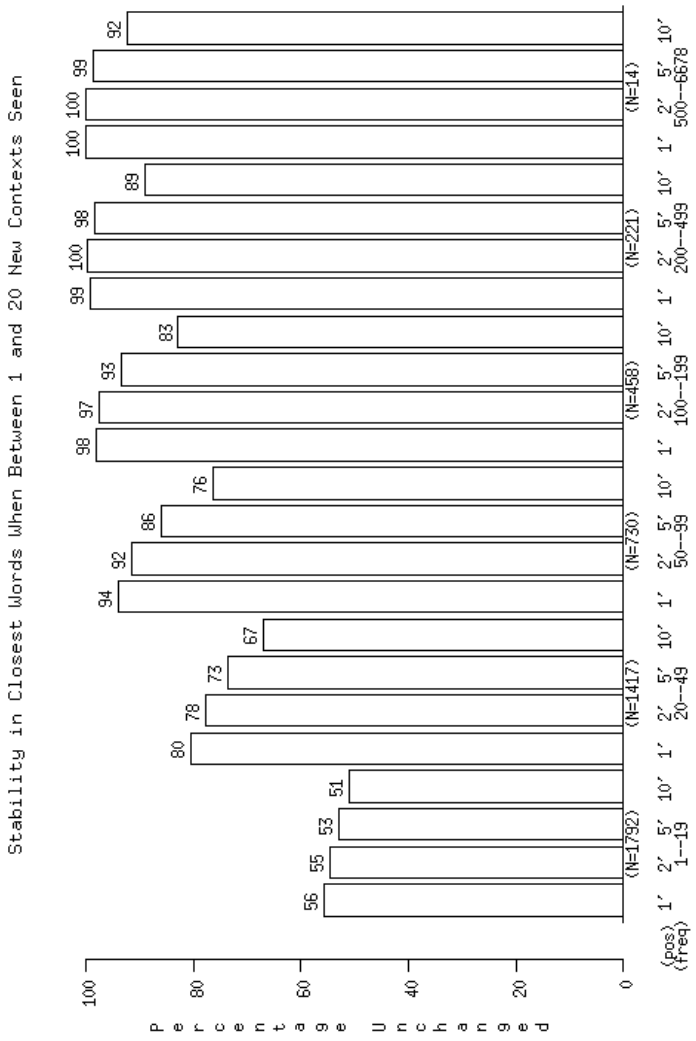

Figure 3.20 Measures of stability in closest words, 1 to 20 new contexts added. Once words possess more than 50 contexts for a word, there is little change in their closest word when 1 to 20 new contexts are added.

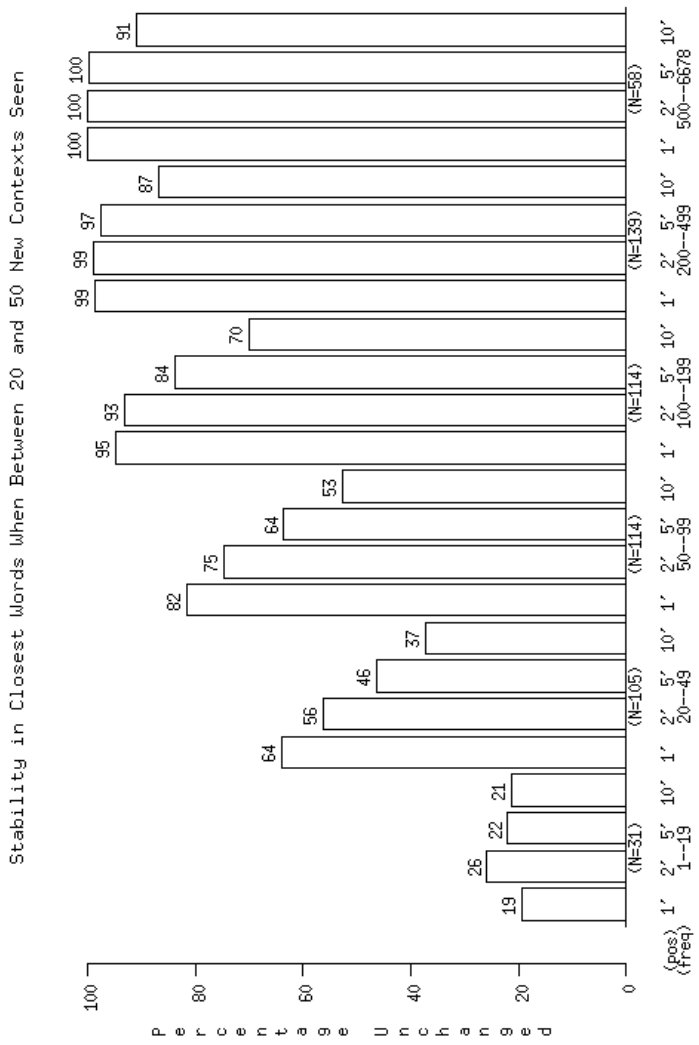

Figure 3.21 Measures of stability in closest words, 20 to 50 new contexts added.

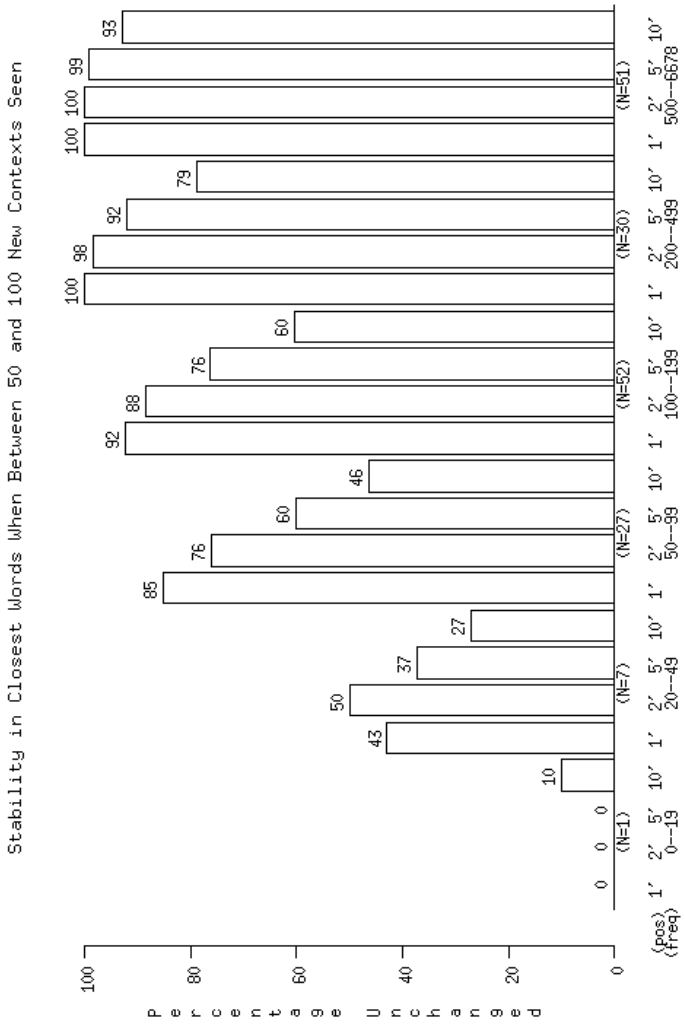

Figure 3.22 Measures of stability in closest words, 50 to 100 new contexts added. As the corpus grows from 1.4 to 2.8 megabytes, doubling the context associated with each word produces fewer and fewer changes.

3.4.3 Basic Vocabulary: The 'Civil Rights Activist' Problem

The work described in this dissertation has been geared to finding similarity between words in a corpus based on their shared syntactic contexts. Individual words have been compared in most of our experimentation.

Yet in any coherent corpus, many concepts are expressed as re-occurring noun phrases of three or more words. In these nouns phrases, there usually exists an internal structure, a binary binding structure common to English and to other noun compounding languages, which joins words in noun phrases in a binary branching dependency tree (Warren 1978; Levi 1978). These structures govern how a long noun phrase is abbreviated. For example, the phrase ''civil rights activist'' can be bracketed as ((*civil rights*) *activist*), which can then be shortened to ''rights activist'', but not to ''civil activist''. On the other hand, a phrase such as ''yale medical library'' is properly bracketed as (*yale* (*medical library*)) which can then be reduced to ''yale library'' or to ''medical library'', but not to ''yale medical''.

Finding the proper decomposition of arbitrary noun phrases is an unsolved problem. It is the problem of finding the basic vocabulary of a domain, that is, the linguistic units (larger than words) in a domain. We are familiar with common examples of such units in everyday language such as *hot dog*. But every corpus defines or redefines such units. For example, in `ADI`, we find the frequently occurring phrase ''information science''. In this phrase, *information* is not being used as an attributive modifying *science*, but rather, *information science* is a unit. As such ''information science curriculum'' should be bracketed as ((*information science*) *curriculum*).

This problem affects SEXTANT's word comparisons in the following way. Without any additional information about basic vocabulary units in the corpus being treated, SEXTANT extracts information from both possible bracketings of a three word noun compound. For example, ''civil rights activist'' yields the following three contexts that enter into the similarity comparison module:

```
right civil
activist right
activist civil
```

The first two items are correct in that they give information about the types of *rights* and the types of *activists* which can be found. But the third context pair is spurious, since although *activists* may possibly be *civil*, this piece of

information should not be inferred from the text. The result of this confusion is that the words *rights* and *activist* would be thought to be more similar than they really are, since both are modified by *civil*, and since similarity is calculated in SEXTANT by counting shared modifiers, in the manner[17] explained earlier.

In early versions of SEXTANT, we saw that such incorrect bracketing had an effect on our output. Individual words were being considered similar because there were appearing frequently in longer noun phrases. The proper way to solve this problem would be to discover the basic vocabulary of the corpus domain, and then to properly bracket within these long noun phrases. We were not eager to tackle this fundamental problem, which must be addressed when units larger than words are considered. Instead we decided to penalize words appearing in such contexts, so that they not be considered similar because of appearing as modifiers in the same noun phrase.

Our method for identifying such contexts was to extract all of those simple noun phrases containing three or more words from the corpus. Then, for each noun phrase, all pairs of modifiers were extracted. For example, the phrase ''private coeducational liberal arts institution'' would penalize any of the following pairs from being considered as similar: *private-coeducational*, *private-liberal, private-arts, coeducational-liberal, coeducational-arts,* and *liberal-arts*[18]. The frequency of any penalized pair can be counted over the corpus. An arbitrary cutoff point was decided; pairs whose frequency was above this threshold were penalized. Our simple penalization technique was to ignore similarity pairs appearing in this list.

3.4.4 Corpus-Based Thesaurus

One of the most potentially valuable aspect of this technique is its possibilities for creating corpus-defined thesauri. As an example, consider the pairs of words found on the following two pages. An *M* in the first column means the word was taken from `MED`, a medical abstract database; an *I* means that it was taken from `CISI`, a database containing documents on libraries and information retrieval. After each word, its frequency in the base appears in brackets, and this is followed by a list of words found to be similar to it in the database. Since the vocabulary of each document collection is slightly

[17] See page 47.

[18] Note that we are not saying that *liberal arts* is not a legitimate noun phrase; rather we are saying that the word *liberal* should not be considered similar to the word *art.*

different and used in different ways, different relations result for the same word.

```
M administration [114] injection | treatment therapy | infusion
I administration [32] graduate office campus | education

M amount [148]        excretion | concentration level activity
I amount [82]         quantity | cost body value | set

M approach [36]       management | intervention | member error
I approach [200]      method | technique model aspect procedure

M aspect [75]         history data symptom management problem
I aspect [177]        approach | model structure | theory

M author [88]         report | problem data case hour response
I author [144]        title journal | paper purpose book report

M cause [75]          etiology | incidence | explanation
I cause [29]          determinant | reason | synonym

M circulation [42]    flow blood-flow | ffa plasma
I circulation [41]    date | profession | supply

M component [77]      antigen | cholesterol constituent
I component [62]      facet | feature design modification

M control [98]        group | child mouse female animal
I control [115]       reference | structure retrieval

M culture [109]       strain bone-marrow line suspension cell
I culture [14]        microstructure | sociologist | virtue

M evaluation [62]     consideration | diagnosis examination
I evaluation [165]    performance analysis | application

M evidence [143]      feature | data finding | case result
I evidence [48]       observation | argument thesis


M failure [46]        reabsorption | resistance | damage
I failure [51]        recall precision | performance usefulness

M feature [87]        evidence | finding pattern significance
I feature [76]        characteristic | component | element factor

M finding [133]       data | feature result observation
I finding [64]        application description contribution

M flow [78]           pressure volume blood-flow plasma | rate
I flow [48]           quantity dissemination | avalanche
```

```
M growth [158]        effect | increase liver hypertrophy tumor
I growth [119]        development interest change rate

M improvement [26]    regression | enrichment similarity
I improvement [60]    effectiveness evaluation availability
```

For example, *administration* was found to be closest to *injection* in the medical database, probably because they both often describe things that are done with drugs. In the library database, one of the closest words to *administration* is *office*.

The word *component* is related to *facet* in the library setting and to *constituent* in the medical one. *Culture* is related to *sociologist* in one field and to *strain* in another. In medicine *improvement* is associated with *regression*, while it is related to *effectiveness* in information retrieval.

These lists demonstrate that even rather common words have rather different meanings and associations in different domains. This is exactly the type of information that a natural language interface must know.

3.4.5 Summary

In this chapter we have presented our two stage technique for semantic extraction from raw text. The first stage employs selective natural language processing, a robust local parsing, that extracts lexical-syntactic contexts for each word in the corpus. The second stage uses a weighted Jaccard similarity measure to compare these contexts and to produce similarity lists for each word. Sample of the results for eighteen different corpora, totaling 50 megabytes of text, are presented in this chapter and in the Appendix. We have demonstrated that this discovery procedure is effective, efficient, and stable. In the next chapter we evaluate the results produced by this technique and show that they do correspond to semantic similarities.

4

EVALUATION

The previous chapter described SEXTANT's partial syntactic extraction technique, and explained how these syntactically derived contexts were used to compare words and produce list of similar words. Visual inspection of these lists gives the intuitive impression that the words on this lists are related. The purpose of this chapter is to demonstrate in some objective manner that the relationships extracted are what are commonly considered as semantic relationships.

We perform this evaluation in three distinct manners. First, we appeal to results from psycho-linguistic literature and show that *free association* results with common adjectives are replicated by SEXTANT. This result may also give some insight as to why humans themselves spontaneously generate certain responses. Second, we create a new evaluation technique for corpus-based linguistics, using what we call *artificial synonyms*. Artificial synonyms are artificially distinguished versions of the same word. We show that, given enough context, these artificial synonyms are recognized as similar by SEXTANT. This technique provides an interesting method of parameterizing how much context is needed to recognize similarity. Third, we measure the performance of SEXTANT against a series of gold standards. These gold standards are human-built general English thesauri and dictionaries. We measure how often SEXTANT is able to reproduce information contained in these sources, and compare our results to those produced by other knowledge-poor techniques.

These results demonstrate that statistical techniques based upon syntactically derived data are able to extract semantically similar words. Chapter 5 shows applications of such information.

The Deese Antonyms

active - passive	alive - dead	back - front
bad - good	big - little	black - white
bottom - top	clean - dirty	cold - hot
dark - light	deep - shallow	dry - wet
easy - hard	empty - full	fast - slow
happy - sad	hard - soft	heavy - light
high - low	large - small	left - right
long - short	narrow - wide	new - old
old - young	rich - poor	pretty - ugly
right - wrong	rough - smooth	short - tall
sour - sweet	strong - weak	thin - thick

Figure 4.1 Deese antonyms. These antonyms are commonly associated in free association experiments.

4.1 DEESE ANTONYMS DISCOVERY

One claim that we are making is that the examination of the lexical-syntactic usage of words over a corpus allows us to extract semantically similar words. To support this claim, we present an experiment in which SEXTANT was able to find exactly and objectively many of the intuitive pairings between semantically similar words in a set studied by Deese (1964).

Psychologists have long used the technique of *free association* as a tool for semantic discovery. The technique consists of presenting a subject with a word, and recording the first word that the subject produces as a response. Many lists of such associations have been collated for use in tests of verbal learning and verbal behavior. Deese collected a list of the most common adjectives and the most frequent responses to them occurring in free association by a large number of subjects. He found that, for the most common adjectives, the most frequently occurring response was a contrastive adjective. These pairs of words, often called the Deese antonyms, appear in the Figure 4.1.

Deese was most interested in finding orthogonal semantic semantics axes which could serve as semantic differentials for other words. Semantic differentials correspond to the unlabeled vectors found in latent semantic indexing discussed in Section 2.4.1. The hope of finding such axes was that word meanings could be plotted within the space of those axes, and meaning

$$\frac{Count(\{Response\ shared\ by\ word_m\ and\ word_n\})}{(Unique\ responses\ to\ object_m\ \times\ Unique\ responses\ to\ object_n)^{1/2}}$$

Figure 4.2 Similarity measure used by Deese to discover semantic axes. The attributes of each word compared were the responses given to that word in free association experiments.

discovered using distance within that space. Deese found that only the groups *big-little-large-small*, *soft-hard-easy-hard*, and *white-black-light-dark* were correlated. In calculating correlation, Deese used the full list of human-supplied free-association responses given by each of 100 subjects as context for each word. A similarity matrix was calculated using the equation given in Figure 4.2, which is similar to the Jaccard coefficient.

Deese contrasted these frequently occurring adjectives with rare adjectives and found (Deese 1962) that uncommon adjectives tend to elicit nouns appearing with that adjective in common noun phrases, rather than contrastive adjectives. For example, the word *administrative* evoked the word *decision* from the noun phrase 'administrative decision.' His hypothesis for this difference was that either common adjectives share a common substitutability in language, or that they correspond somehow to ''some natural perceptual or cognitive property.'' He concluded that both possibilities probably play a role in associative meaning. Justeson & Katz (1991) hypothesize another factor probably contributing to these associations. They showed, by considering only marked adjectives and appearance in the same sentence, that over the tagged BROWN corpus, antonymous adjectives tend to occur significantly more often together in the same sentence than chance would dictate.

SEXTANT provides a different sort of data, less subjective than Deese's human supplied association lists, and comparable to Justeson's and Katz's data on co-occurrence in the same sentence. Finer grained, empirical data on word use is supplied by the parsing mechanism of SEXTANT. We can see how each word is modified or what each modifies. In order to examine what this information tell us for the Deese antonyms, we performed the following experiment on the 6 MB SPORTS corpus, described in the Appendix. The corpus was parsed as described in Chapter 3, and SEXTANT compared the modifying words among themselves, using what they modified as attributes. Each modifier was compared with each of the 14,000 other unique modifiers appearing more than once in the parsed corpus. The closest words to each of the following Deese antonyms are given in Figure 4.3a.

Modifier [*Contexts*]	*Groups of similar modifiers*
large [844]	SMALL | important major | great various main different field new
small [725]	LARGE | major | field new important various area time | state
new [667]	major different state | diverse modern early field various time
high [492]	LOW | average different | time level increase major area point
long [353]	SHORT | time single vertical surface cold small front different
light [259]	HEAVY different wave energy surface particle | wind magnetic
strong [232]	WEAK | different primary movement natural good important work
black [212]	WHITE | woman lead successful major | top school chinese popular
short [202]	LONG | prose jump entire slender | distance run powerful classic
heavy [188]	LIGHT | total solid gas excessive weight body difficult annual
low [175]	HIGH | increase | average | air oxygen ocean temperature excess
deep [160]	SHALLOW | warm surround subtropical coastal | cold depth
young [153]	popular able lead | hunter OLD famous age school numerous
white [151]	BLACK red | blue brown dark | hard head color gray top
shallow [148]	DEEP | temperate inland pressure warm | cloud freshwater
wide [143]	NARROW | worldwide widespread color matter | broad international
right [136]	amendment LEFT privilege freedom | constitution force equal
dry [132]	WET | moist cold | surround warm region layer natural ocean
little [128]	considerable | way body sufficient temperature additional
old [127]	country traditional similar present place NEW | indian original
cold [122]	warm | HOT | coastal | warmer moist temperature temperate
active [119]	principal movement system maximum behavior return instrument
thin [118]	THICK | upper transparent bony gas wood waxy exterior | metal
good [105]	poor | excellent change live | strong active different natural
hot [100]	warm COLD | pure cooler warmer solution | geothermal molten mine
full [87]	congress education | increase maximum day constitution party
hard [84]	solid | SOFT | leathery | sedimentary shape layer exterior liquid
poor [84]	good | adequate | moist real local health child student respect
rich [83]	warm vast coastal surround mediterranean region | distinct
dark [69]	bright side | eye horizontal brown red planet visible white
smooth [68]	plane concave convex | porous fracture quiet thin silicate
top [68]	nhl clay rocky football team horizontal outstand | magnificent
front [67]	hind | tube absent horizontal antenna tail leg cylindrical
narrow [62]	WIDE edge | tidal floor chamber rectangular row main bottom rear
bottom [58]	zone | layer underground gill nile pipe allow mineral | surrounding
rough [58]	damp pure outer exterior brick phase tube | loose ceramic saline
soft [55]	HARD loose | solid surround alkaline | dead molten bare
weak [55]	STRONG | motion distinct mercury electromagnetic perpendicular
left [51]	hind RIGHT | follow beat front tennis oppose elbow opponent
slow [47]	rapid reaction fluid continuous | electron object inorganic
dead [44]	host prior snake resort careful sting | aquatic soft nerve cheek
sweet [43]	important

Figure 4.3a Closest words to Deese Antonyms for frequently occurring adjectives.

Modifier [*Contexts*]	*Groups of similar modifiers*
wet [39]	DRY moist | damp quartz favorable warm | bonding sticky
big [38]	modest asian confidence hog paddle combination strategic link
thick [36]	THIN outermost | oxide outer | cement hazardous brownish stiff
back [35]	conical | sharp brick throat | reflective take quarry initial
clean [21]	cooler | potable vicksburg amputation | etch bottled | surplus
easy [19]	persia libya expressionism telephone optics | prime dipolar
bad [17]	decree | number-one | doubt foul favorable tile | major-league
tall [16]	iconostasis | courtyard traveler | tower | exterior axis
fast [14]	imaginary | harvest | indoor long-distance high-velocity
empty [13]	exhibition | row architecture heat blue | dark urban wide
inside [9]	absorb thin-wall | millimeter jaw epithelial | sac lining
passive [8]	establishment | discovery mirror | influential earth society
dirty [6]	whipple | subsoil | darkness bulb | archaic | tap cap ether
happy [4]	decade ordinary | broad level coastal domestic lake upper
sour [4]	court water
wrong [3]	part single play

Figure 4.3b Closest words to Deese Antonyms for less frequently occurring adjectives.

Some of the Deese antonyms pairs were not represented in the data given to SEXTANT, either because one element of the pair was not returned by the morphological analyzer as a noun or an adjective (e.g., *alive-dead*), or because the word was not present in the corpus (e.g., *happy-sad, pretty-ugly*).

Of the remaining 30 pairs, the Deese response was found as the first or second closest word 14 times. Consider that, for each word, similarity is calculated to all of the other 14,000 unique modifiers appearing more than once in the corpus. These pairs are *black-white, cold-hot, deep-shallow, dry-wet, hard-soft, heavy-light, high-low, large-small, left-right, long-short, narrow-wide, strong-weak,* and *thick-thin*. These pairs are also the pairs for which we have the most context, as can be seen in the above list sorted on frequency of appearance of the word in the corpus.

In a few other cases, words were paired with synonyms to the Deese pairings. *Dark* was matched to *bright* rather than *light*; *slow* was matched to *rapid* rather than *fast*. *Front* was found most similar to *hind* rather *back*. This last case is due to the corpus bias. Since it consists of sentences concerning sports and sport-related animals, *front* and *hind* are often found modifying *foot, limb, leg* and *toe*.

Some words have bizarre associations, such as *tall-iconostasis* and *top-nhl*. *Iconostasis* only appears three times as a modifier, and modifies *altar, screen, clergy*. *Tall* only appears 16 times as a modifier, and also modifies *altar* and *screen* in different sentences. The only word in common with *short* and *tall*, the Deese association pair, is *leg*, which occurs once in this corpus. But since *short* also modifies 99 other unique words, *tall* and *short* are not seen as close by the Jaccard measure. This is a problem with the technique whenever a word does not have enough context with which to judge its similarity.

As for the pair *top-nhl*, they share the following attributes: *defenseman, player,* and *team*. But one of these shared associations stems from the phrase *nhl top defenseman*, creating the spurious effect discussed in Section 3.4.3. Since *nhl* and *defenseman* appear so infrequently in the corpus, this effect dominates the similarity calculation. *Top* and *bottom* share the four words *surface, water, line* and *layer*, but these words appear often in the corpus and have low weights[1] giving rise to the seemingly bizarre associations mentioned in the preceding paragraph.

These results show two things. First, they support the claim that SEXTANT uses a context for words that permits the extraction of semantically similar words, since many of the Deese antonyms are found to be maximally similar to each other over a previously untagged corpus.

Second, they help to explain the different association behaviors of frequently occurring and rare adjectives that Deese observed. As stated above, rare adjectives are associated with nouns with which they commonly appear, and common adjectives are associated with other adjectives. In this experiment, SEXTANT compares adjectives and adjectively used nouns by considering all the words that they modify. Since these adjectives and nouns are considered similar by SEXTANT when they modify many of the same nouns, and since SEXTANT brings the Deese antonyms together as most similar, this means that these common adjectives modify many of the same nouns. With rare adjectives few nouns are involved in these modifier-modified pairs. This observation leads to the hypothesis that the cognitive load of associating a noun to a common adjective, since there are so many from which to choose, is heavier than that of associating a noun to a rare adjective, where the choice is limited. For a common adjective, since many nouns are involved, it would seem that the subject falls back on the other adjectives that modify the nouns associated with the common adjective. This linking then activates those

[1] See Section 3.2.6 for attribute weighting.

adjectives, i.e., the Deese antonyms, which occur most frequently with these nouns.

4.2 ARTIFICIAL SYNONYMS

In order to evaluate a sense disambiguation method, Schutze (1992) introduces the idea of artificial ambiguous words, words which are morphologically distinct but which are considered as identical in the same way that homographs are identical in text. Schutze is interested in using windows of words within 1000 characters of a given word to provide a disambiguation context. To test his methods, he creates a number of artificial ambiguous words (*author/baby*, *giants/politicians*, and *train/tennis*) and makes his program consider them as identical strings. He then performs discriminant analysis on the context of each string to separate the two senses of his artificial words.

Borrowing this idea with a slight twist, we define and use the idea of *artificial synonym* in this section. A synonym is defined in Webster's as a ''one of two words . . . that have the same essential meaning.'' We create artificial synonyms by altering the same word to appear as two different strings throughout the corpus. Since each string is really the same word the two different strings are the purest form of synonym, such as occurs naturally in national spelling variations, e.g., *tumor-tumour*.

Since synonyms are the most similar words and SEXTANT is looking for similar words, we can use this technique to calibrate our system's powers and limits. For example, we can replace half of the occurrences of a word such as *cell* by the word *CELL*. These two strings are considered as two different words by SEXTANT which maintains case distinctions. If SEXTANT can discover their similarity among all the other words in the corpus, this supports our claim that lexical-syntactic context can be used to recognize similarity for highly synonymous words.

Similarly this technique can be used to measure how much context SEXTANT needs to recognize similarity, by iteratively replacing different percentages of a word's appearance in the corpus by its artificial synonym. We can measure the point at which similarity is no longer recognized.

4.2.1 Experimentation

According to the Zipf distribution of word usage throughout any corpus (Zipf 1965), a small number of words are used very frequently, a slightly larger number are used a little less frequently, and the vast majority of words are used very rarely, making a graph of frequency-to-rarity that drops off very sharply. Similarly, in all of our corpora, the quantity of context by which we can judge a word's meaning follows the same steep slope. We classify the words of a corpus into four groups: Frequent (the top 1% of words in order of frequency), Common (the next 5%), Ordinary (the next 25%), and Rare (all the rest). In the MED corpus, whose plot is shown in Figure 4.4, the Frequent words accounted for 26% of the data by which SEXTANT judges similarity, the Common words for 33%, the Ordinary words for 29% and the Rare words for 11%.

We used the MED corpus, which possesses 5900 unique nouns to be compared for our experiment. We randomly chose 20 words from each category, and for each word ran the following experiment.

> **Procedure 1.** Randomly extract F percent of the context pairs including word w from the context pairs for the entire corpus. Alter w in these pairs. This alteration is done by transliterating w to uppercase since all the data is usually in lower case. Reintroduce the altered context pairs into the corpus data. Calculate the similarity of all words. Measure the position at which the altered word appears on the original word's similarity list, and vice versa. If the altered word is at position 1, then it has been recognized as most similar to the original word.
>
> Perform the experiment for each word with F = 50, 40, 30, 20, 10, 5, and 1.

4.2.2 An example: *patient-PATIENT*

When we ran this experiment over the MED corpus, one of the Frequent words randomly extracted was *patient*. Over the MED corpus, SEXTANT extracted 886 contexts for *patient*. At the F = 50 level, these contexts were extracted from MED data. Of these pairs, 430 were then randomly extracted and the first word transliterated to upper case, giving the type of data seen in Figure 4.5.

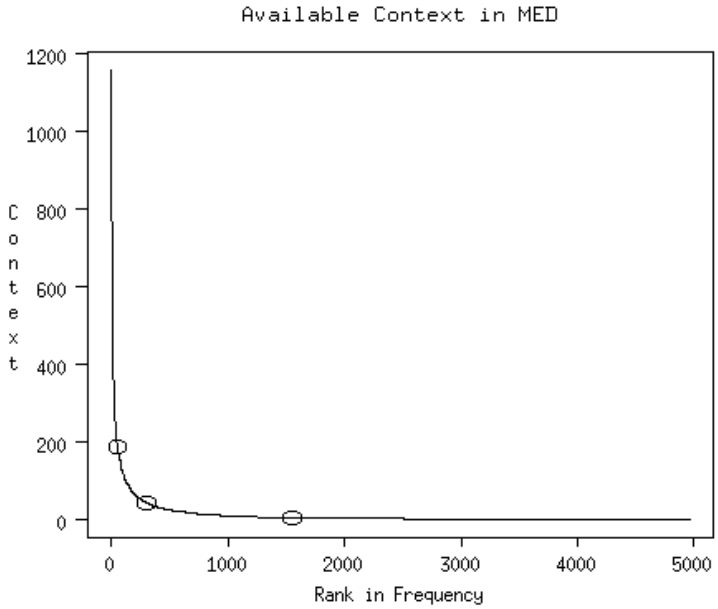

Figure 4.4 Number of attributes available for the nouns in the MED corpus.The words are presented from the most frequent to least, and the y-axis gives the number of attributes for each word. Few words have many attributes, and most words have very few, following the Zipf distribution of words in text. Left of the first ball are the Frequent words, to the left of the second ball are the Common words, to the left of the third ball are the Ordinary words. The Rare words are to the right of the third ball.

. . .	. . .
patient proximal	PATIENT treat-DOBJ
patient procedure	PATIENT discontinue-DOBJ
patient responsive	PATIENT arthritis
patient indicate-DOBJ	PATIENT describe-DOBJ
patient cancer	PATIENT consider-DOBJ
patient series	PATIENT reflect-SUBJ
patient day	PATIENT suspect-IOBJ
patient non-hemophilic	PATIENT syndrome
patient control	PATIENT adult
. . .	. . .

Figure 4.5 Attributes of original word and its artificial synonym.

patient [457]	PATIENT | case | group | child day treatment woman
PATIENT [429]	patient | case | child group | treatment study result

Figure 4.6 Most similar words to *PATIENT* and *patient* when their context is evenly divided between the two word forms.

patient [799]	case | group child | treatment PATIENT result day
PATIENT [87]	patient woman | incidence site child year diagnosis

Figure 4.7 Similarity results when *PATIENT* only has one-tenth the context of *patient*.

patient [832]	case | group child | treatment result study day
PATIENT [54]	iv instance benefit induction curve woman patient

Figure 4.8 Similarity results when *PATIENT* only has one-twentieth the context of *patient*.

Then the following three groups of data were rejoined: (1) the MED data with no *patient* data, (2) the 457 nontransliterated *patient* contexts, and (3) the 429 transliterated *patient* contexts. Next the similarity calculations were run over this combined data, corresponding to the original data but with 429 incidences of the word *patient* appearing as *PATIENT*. The result of the similarity calculations for each word given in Figure 4.6 show that the artificial synonym *PATIENT* was recognized as the most similar word to *patient* at the 50% level, and vice versa.

Running the same experiment of 40% of the context of *patient* changed to *PATIENT* gave the same result, as did 30%, and 20%. At the 10% level, see Figure 4.7, *PATIENT* is represented by only 87 context words, and begins to drift away from *patient*, though *patient* is still recognized as the closest word to *PATIENT*.

At the 5% level, shown in Figure 4.8 832 of the *patient* contexts remained in lower case, while 54 were transliterated, and *PATIENT* becomes the 21st closest word to *patient*, while *patient* is the 7th closest word to *PATIENT*.

At the 1% level, only 13 of the 886 contexts were transliterated to *PATIENT* and the artificial synonym falls to 431st closest of the 5358 words considered, while *patient* is ranked as the 34th closest word to *PATIENT*.

FINDING AN ARTIFICIAL SYNONYM							
	Percentage of context changed into Artificial Synonym						
	50	40	30	20	10	5	1
Frequent (N = 889 . . . 204)							
Art. Syn. 1st or 2nd most similar	100	100	100	90	35	5	
3rd to 5th				10	15	5	
6th to 10th					10	10	
Farther than 10th					40	80	100
Common (N = 180 . . . 49)							
Art. Syn. 1st or 2nd most similar	80	65	65	50	10	5	
3rd to 5th	10	20	10	10	15		
6th to 10th		5		10	20		
Farther than 10th	10	10	25	30	55	95	100
Ordinary (N = 42 . . . 6)							
Art. Syn. 1st or 2nd most similar	25	5	15	10			
3rd to 5th	10	15		5			
6th to 10th							
Farther than 10th	65	80	85	85	100	100	100

Figure 4.9 Given an original word, find the artificially created synonym.Twenty words were randomly chosen from each frequency group. In function of the percentage of incidences changed into an artificial synonym, the position of the artificial synonym in the original word's similarity list is given.

4.2.3 Results

As stated above, this experiment was run for 20 randomly selected words from each of the Frequent, Common, and Ordinary classes of words. Figure 4.9 presents the results of finding the Artificial Synonym from the point of view of the original synonym. For example, when 10% of the incidences of the word *patient* were changed to *PATIENT*, this artificial synonym was ranked as the 5th closest word to the original *patient*. Since *patient* was one of the 20 Frequent words extracted, this result counts for one-third of the 15% appearing in the 10% column of the row *Frequent, 3rd to 5th* table.

Figure 4.10 gives the results from the point of view of the artificial synonym and shows how the original word is judged regarding similarity as the context share of the artificial synonym decreases. Recall that in example given page 78 where *PATIENT* took 10% of the context from the word *patient*, the original *patient* was still the most similar word even though *PATIENT* only had 87 context words and *patient* retained 799. This particular case accounts for one-seventh (1/7) of the 35% under the 90% column of the row *Frequent, Orig. Syn. Most or 2nd most similar* Figure 4.10.

FINDING ORIGINAL SYNONYM							
	Percentage of context retained by Original Synonym						
	50	60	70	80	90	95	99
Frequent (N = 889 . . . 204)							
Orig. Syn. 1st or 2nd most similar	100	100	100	100	35	5	25
3rd to 5th					15	25	10
6th to 10th					15	5	
Farther than 10th					35	65	65
Common (N = 180 . . . 49)							
Orig. Syn. 1st or 2nd most similar	70	70	75	40	20	20	
3rd to 5th	15	10		20	25	15	
6th to 10th	5			5	20	10	
Farther than 10th	10	20	25	35	35	55	100
Ordinary (N = 42 . . . 6)							
Orig. Syn. 1st or 2nd most similar	30	20	10	10			
3rd to 5th	5		5				
6th to 10th				5			
Farther than 10th	65	80	85	85	100	100	100

Figure 4.10 Finding the original word. Given an artificially created synonym, shows the frequency with which the original word appears in certain positions of the artificial synonym's similarity list, in function of the percentage of context retained by the original word.

4.2.4 Discussion

The above results show that the more context that one has to use in making judgements, the better the results. The Frequent words have hundreds of context points by which to judge them. Taking only 10% of these contexts points creates an artificial image of the original word which often is recognized within the five most similar words (out of thousands of candidates). The Common words (each with 50 to 200 context points) also perform well down to a 20% transformation level, at which their artificial synonym receives only 10 to 40 context points.

The surprising result of these experiments is the asymmetry that can be seen in the two tables. In the first table, as the context of the original word grows larger than that of the artificial synonym, say as the context becomes 10 times as great, the artificial synonym gets swamped by other words possessing more context, and the matches in context between the original word and the artificial synonym become less important in the Jaccard measure. This is so because the original word possesses so much context that (a) other frequently appearing words might share significantly more with it than does the artificial synonym,

and (b) the number of attributes that the original and artificial synonym do not share becomes a dominating factor in the denominator of the Jaccard measure.

However, in the list of words most similar to the artificial synonym, the original word remains highly similar much longer as the context of the artificial synonym dwindles. As Figure 4.10 shows, even as the number of context points for the artificial synonyms melts away to a few dozen (at the 90% unchanged level for Frequent words; at the 80% unchanged level for Common words) the similarity of the original word is still recognized among the thousands of candidates. Although the number of attributes not shared by the original and artificial synonym is the same as in point (b) in the previous paragraph, the rare artificial synonym seems to match more closely with the original word than with others in the corpus.

If this effect of synonymy can be extended to truly synonymous words, a rule of thumb to retain here might be: *We can give greater credence to matches of rare words with common words than we can give to matches of common words with rare words.*

4.3 GOLD STANDARDS EVALUATIONS

As a further evaluation of our similarity extraction techniques, we measure our results against a series of gold standards. By gold standard, we mean some human-compiled collection of related words. Such collections appear in ordinary thesauri, the most famous of which is *Roget's Thesaurus*. A version of *Roget's Thesaurus* (the 1911 edition) is freely available on the Internet[2]. In it are collected more than 30,000 unique words under 1000 topic numbers. Also, we had available to us[3] another, more recent, general English thesaurus developed at the University of Macquarie in Australia.

We also use an online dictionary, *Webster's 7th Edition*. Many researchers have drawn on online dictionaries in attempts to do semantic discovery, as we noted in Section 2.3. In our work, the thesauri and dictionary are used only as a tool to evaluate the information extracted by SEXTANT and to demonstrate that SEXTANT's results do overlap with these manually created sources.

[2] For example, in March 1993 it was available via anonymous ftp at the Internet site *world.std.com* in the directory /obi/obi2/Gutenberg/etext91, as well at over 30 other sites listed by the ftp server *archie*.

[3] In work done at the Laboratory for Computational Linguistics, Carnegie Mellon University, director Prof. David Evans.

The premise of this testing is, since we claim to be recognizing semantically similar words, that some of the relations that we discover should appear in analogous lists that have been made by hand. This will not always be the case, since the relations that we extract are corpus dependent and may include some relations which are not expressed in a work intending to describe relations in general English. An example of this is the similarity relation that we find in the MED corpus between *injection* and *administration*. When we look up *administration* in Webster's we see:

> **ad-min-is-tra-tion** n. 1. the act or process of administering 2. performance of executive duties :: c<MANAGEMENT> 3. the execution of public affairs as distinguished from policy making 4. a) a body of persons who administer b) i<cap> :: a group constituting the political executive in a presidential government c) a governmental agency or board 5. the term of office of an administrative officer, or body.

For *injection* we get a definition which has no words in common with *administration*:

> **in-jec-tion** n. 1. an act or instance of injecting (as by a syringe or pump) 2. something (as a medication) that is injected.

In Roget's, *injection* and *administration* appear under distinct topic headings. *Injection* appears in topic 300 (Forcible ingress: Insertion) and topic 662 (Remedy). *Administration* is found in topics 693 (Direction), 737 (Authority), and 786 (Apportionment).

The same problem of generality of meaning appears in the newer Macquarie thesaurus, which lists *injection* with other remedies and ingresses, and *administration* with taxes, bureaucracies functions, management, and allotments.

In this case, it might seem that since *administration* and *injection* are nominalized forms of verbs, one should search for the verbs *administer* and *inject* in these gold standard sources. But *inject* and *administer* suffer the same orthogonality in these sources as their nominalized forms.

We are not suggesting that works such as Roget's or Webster's are not respectable and authoritative sources of English usage, but rather that word usage in subdomains of English is not necessarily represented in these works. Despite these drawbacks, gold standards provide synonyms for numerous words, and we can measure how many times that the similarities that SEXTANT finds are reflected in them.

4.3.1 Roget and Macquarie

Both of these thesauri arrange words in shallow hierarchies of specific to general terms. *Roget's Thesaurus* defines 1000 headings ranging from Existence (Topic Number 1), Inexistence (2), Substantiality (3), Unsubstantiality (4), up to Rite (998), Canonicals (999), and Temple (1000). The Macquarie thesaurus is significantly more modern with a wider coverage. It possesses 824 main headings such as Abstinence, Accounting, Accusation, Achievement, Addition, . . . , Woman, Word, Work, and Worker. These headings are further divided into more than 12,000 specific subheadings. For example, Vehicle is divided into the subheadings: vehicle - car - carriage - wagon - truck - tram - train - bicycle - pram - sledge - miscellaneous vehicles - vehicle parts - vehicular.

We can use these thesauri as a reflecting glass for SEXTANT results. We have been claiming that SEXTANT's use of syntactic contexts allows discovery of similar words, just as humans deduce the meaning of a new word by comparing contexts. When SEXTANT finds two words to be similar, we can look in the gold standard and see if they have placed SEXTANT's 'similar words' under the same heading or subheading.

4.3.2 Method

Separately then, from Roget's thesaurus and from the Macquarie thesaurus, we extracted each single-word entry, since our research has been examining similarity at the word level. From Roget's, we extracted 60,071 individual words and stored each word with its topic number or numbers. A portion of the extracted Roget list in Figure 4.11 shows that *abatement* appears under two topics: Nonincrease (36) and Discount (813). *Abbe* and *abbess* both belong under the same topic heading 996 (Clergy).

The extracted Roget's list then has about 60,000 words (an average of 60 words for each of the 1000 topics). Of these 32,000 are unique (an average of two occurrence for each word). Assuming that each word appears under exactly 2 of the 1000 topics, and that the words are uniformly distributed, the chance that two words w_1 and w_2 occur under the same topic is

$$P_{Roget} = 1 - (998/1000)^2,$$

since w_1 is under 2 topic headings and since the chance that w_2 is under any specific topic heading is $2/1000$, or about 0.4%.

Roget's	
entry	*Topic*
...	
abatement	36
abatement	813
abatis	717
abatjour	260
abattis	717
abattoir	361
abba	166
abbacy	995
abbatial	995
abbatical	995
abbatis	717
abbe	996
abbess	996
...	

Macquarie	
entry	*subheading*
...	
disesteem	036406
disesteem	063701
diseur	022701
disfavour	003901
disfavour	056601
disfavour	063701
disfeature	018212
disfeaturement	018201
disfigure	006804
disfigure	018212
disfigure	020103
disfigured	006803
disfigured	020102
...	

Figure 4.11 Samples from One Word Entries in Both Thesauri

From the Macquarie thesaurus, the 130,675 one-word entries were extracted with their subheading number. There were 5602 unique subheadings. As seen in Figure 4.11, the word *disesteem* is found under the heading Low Regard(0364) and subheading Hold In Low Regard(06) as well as in Disrepute(0637) subheading Disrepute(01). *Disfavour*[4] is also found in this last subheading (063701). In the Macquarie extraction, a one word-term appeared under an average of 2.12 subheadings. Assuming that each word appears under 3 subheadings, and that the words are uniformly distributed, the probability that 2 words w_1 and w_2 appear under the same subheading is

$$P_{Macq} = 1 - (5599/5602)^3,$$

since w_1 is under 3 subheadings and since the chance that w_2 is under any specific subheading is 3/5602. The probability, then, is about 0.16%.

4.3.3 Evaluation Experiment

Though there are some drawbacks with using general language thesauri on results from specific corpora, we decided to perform the following experiment.

[4]Note the English spellings of words in this Australian-built thesaurus have not been changed.

> **Procedure 2.** Given a corpus, use SEXTANT to derive similarity judgements between the nouns appearing in the corpus. For each noun, take the noun appearing as most similar. Examine the human compiled thesaurus to see if that pair of words appears under the same topic number or subheading. If it does, count this as a hit.

We chose three corpora: HARVARD, derived from *Grolier's* and focussed on institutions; SPORTS, also derived from *Grolier's* but dealing with sports; and MERGERS, from the *Wall Street Journal*[5]. These corpora are sufficiently coherent for SEXTANT to produce good results, while touching on concepts likely to be included in our two general English thesauri.

SEXTANT was run over each of these corpora and similarity lists produced. Each word was paired with its most similar word, and all the pairs from each corpus were sorted according to the corpus frequency of the first word. We would expect the words having the most context to produce the best similarity relations. Each pair was looked up in the previously derived *Roget's* list, then in the Macquarie list, then in both combined. The existence of at least one hit was counted for each pair. For example, matching the top 20 closest pairs found by SEXTANT from the MERGERS corpus against the *Roget's* list generated the hits shown in Figure 4.12.

The pair of most similar words is followed by the number of context points, for the first word of the pair, available in the corpus. The fourth column lists topic headings and topic numbers under which both members of the pair appear in Roget's. For example, *company* appears modified by 7625 other words in the corpus, and is found most similar by SEXTANT to *concern*, but these two words do not appear together under any heading in *Roget's*. *Offer* appears modified by 2744 other words, is found closest to *bid*, and both *offer* and *bid* appear under topic number 763 (Offer) of *Roget's*.

Of the first 20 words (ranked by frequency) of MERGERS, 8 hit in *Roget's*, making a 40% hit rate (see comprehensive table in Figure 4.14.

Since each of these three corpora are coherent, one might expect that any randomly chosen pair of frequently occurring words would hit in one of the thesauri. To obtain an experimental measure of this effect, we created random pairs by extracting nouns randomly from the data file of noun-attribute pairs used to make the similarity judgements for the SPORTS corpus. Since the frequency of a noun in that file equals the number of its contexts, the

[5] See the Appendix for details on each of these corpora.

discovered word pair		*frequency 1st word*	*Roget Topic Match*
company	concern	7625	---
share	stock	6241	---
stock	share	2801	---
offer	bid	2744	--- Offer(763)
stake	share	2661	---
business	operation	2643	--- Action(680)
unit	subsidiary	2623	---
sale	transaction	2609	---
bid	offer	2504	--- Offer(763)
year	price	2350	---
group	company	2316	--- Assemblage(72)
price	value	2295	--- Goodness(648), Price(812)
analyst	group	2174	---
plan	proposal	2022	--- Plan(626)
executive	official	2014	--- Authority(737)
market	business	2003	---
concern	company	1870	---
bank	group	1850	---
agreement	plan	1738	---
firm	concern	1592	--- Merchant(797)

Figure 4.12 First twenty pairs discovered in MERGERS, and their overlap with *Roget's*

discovered word pair		*random word pair*	
water	field	water	record
field	surface	field	order
surface	field	surface	tribe
court	supreme	search	court
play	work	play	color
role	part	role	court
game	sport	game	master
system	field	system	guild
form	type	form	academy
player	game	player	orbit
area	region	area	law
part	role	part	actor
work	play	work	performer
skill	development	skill	imbalance

Figure 4.13 Discovered and Random Pairs from SPORTS

Percentage of hits in Hand-Made Thesauri

	HARVARD			SPORTS			MERGERS			RANDOM		
rank	R	M	E	R	M	E	R	M	E	R	M	E
1-20	50	40	50	55	70	70	40	55	55	5	0	5
21-40	35	45	50	40	45	50	25	25	35	10	10	15
41-60	30	35	50	30	35	50	25	35	40	0	0	0
61-80	35	40	45	25	40	50	10	25	30	5	5	5
81-100	35	30	35	40	40	50	10	20	25	5	0	5
101-200	30	34	44	36	29	46	23	24	34	1	5	6
201-300	31	29	39	34	39	51	20	20	31	3	2	4
301-400	17	20	26	18	23	30	20	23	31	3	3	5
401-500	14	13	20	34	36	46	7	14	16	7	3	8
501-600	12	13	18	19	21	30	10	10	18	2	4	5
601-700	10	15	20	23	25	35	8	8	12	3	2	5
701-800	11	9	16	17	21	28	10	15	18	0	1	1
801-900	6	7	11	21	21	25	10	13	17	2	2	4
901-1000	8	8	14	11	21	24	2	4	5	4	5	7
1001-2000	5	5	8	7	10	13	2	2	4	2	2	4
2001-3000	3	3	5	4	6	8	1	2	3	1	2	3
3001-4000	3	3	4	2	4	5	1	1	2	2	2	3
4001-5000	2	2	3	3	4	6	1	1	2	0	2	2
5001-6000	1	1	2	2	3	5	1	1	2	0	1	1
6001-7000				2	3	5				1	1	2
7001-8000				1	2	4				1	1	1
8001-9000				1	3	4				0	1	1
9001-10000				1	2	3				0	1	1

Figure 4.14 Table of hits into *Roget's* (R), *Macquarie's* (M), or either (E) Thesauri.

probability of choosing a word is proportional to that frequency. Randomly chosen words were paired with the most frequently occurring words, and the hit rate of these randomly formed pairs was calculated. These hits rates (in the Roget's, Macquarie's and combined thesauri) of the random pairs from the SPORTS corpus are shown in the fourth column set of Figure 4.14.

4.3.4 Analysis of Results

From the table in Figure 4.14 it can be seen that many frequently occurring words hit in at least one of the two hand-built thesauri. For the HARVARD corpus, 47 of the first hundred pairs hit; for SPORTS, 54 hit in at least one; and for MERGERS, a less general corpus, 37 of 100 hit. Of the random pairs, only 6 of the 100 most frequent words was paired with a word that results in a hit in a thesaurus.

The first 20 pairs from HARVARD scored hits under the *Roget* topics: Power, Production, Teaching, School, Authority, Religious Knowledge, and Piety. Surprisingly, pairs from this group identified as similar by SEXTANT which did not appear together under any Roget topic were *church-school, settlement-institution, constitution-government, group-institution, work-school* and *state-law*. The reasons for the absence of hits in cases like these can be one of the following:

1. One of the two words in the discovered pair does not exist in the hand-coded source. We counted the number of times that only the first element of the pair did not appear in each thesaurus, since one might expect if the first one did appear that SEXTANT should be able to find some similar word that also appeared. The absence of the first element of a discovered pair accounted for 40%, 43%, and 46% respectively of the misses in *Roget's* and for 32%, 28%, and 37% respectively of the misses in *Macquarie* for the HARVARD, SPORTS, and MERGERS corpora. Excluding these pairs from the table presented in Figure 4.14 improves those words ranked greater than 600 by only a few percentage points.

2. The similarity between the words in the discovered pair is semantically grounded (e.g., *company-concern*, *movement-group*), and present within the corpus, but the words simply never appear together in the thesaurus. This can be due to the fact that the thesaurus is a general English thesaurus, and that the axis along which the similarity exists is too technical or domain-specific to be included. For example, the business sense of *concern* did appear in the 1911 *Roget's*[6], while that of *company* did not.

3. The words in the pair were similar but along an axis of hyperonymy, (e.g *student-child*, *organization-agency*, *doctrine-idea*), or meronymy (e.g., *century-year*, *leader-group*).

4. The words in the pair had nothing or little in common, and a limitation in SEXTANT's power brought them together (e.g., *state-law*, *center-development*, *number-education*).

These last three categories blend into one another in a smooth fashion, and missing word pairs cannot be classified precisely as belonging exclusively to one or another. At this point, subjective judgments as to whether a pair should

[6] Unfortunately this old edition of *Roget's* is the only one available via anonymous ftp. Other electronic editions of Roget's do exist, and a newer edition may alter the numbers given in Figure 4.14 without changing the overall result.

have been included, similar to the ones made in the construction of the original thesaurus, must be made.

As another angle on verifying relatedness, we evaluated our results using *Webster's 7th* as an alternative measure. Researchers such as Plate (Wilks *et al.* 1989) and Sparck Jones 1986, have placed great store in using machine readable dictionaries as a resource for computational semantics, as we have noted in section 2.3. They used these sources by reducing dictionary entries to a list of individual words, eliminating stopwords. The remaining words were considered as independent semantic tokens describing the head word of the entry. In the next section we use similar techniques with *Webster's 7th* to further verify that SEXTANT does extract some degree of similarity.

4.4 WEBSTER'S 7TH

Our use of *Webster's 7th* in evaluating discovered similarity pairs is based on the assumption that similar words share some overlap in their dictionary definitions. In order to determine overlap, each entire literal definition is broken into a list of individual words. This list of tokens contains all the words in the dictionary entry, including dictionary-related markings and abbreviations. In order to clean this list of non-information-bearing words, we automatically removed any word or token

1. of fewer than 4 characters,
2. among the most common 50 words of 4 or more letters in the Brown corpus,
3. among the most common 50 words of 4 or more letters appearing in the definitions of *Webster's 7th*,
4. listed as a preposition, quantifier, or determiner in CLARIT's lexicon,
5. of 4 or more letters from SMART's stoplist,
6. among the dictionary-related set: *slang*, *attrib*, *kind*, *word*, *brit*, *-ness*, *-tion*, *-ment*.

These conditions generated a list of 434 stopwords of 4 or more characters; they are listed in the appendix on page 151.

administer, administering, administrative, affairs, agency, board, constituting, distinguished, duties, execution, executive, government, governmental, making, management, office, officer, performance, persons, policy, political, presidential, public, term

Figure 4.15 List extracted from *Webster* definition of ''administration,'' after removal of short words and filtering through stoplist.

similarity pair	*freq*	*intersection*
company - concern	7625	organization
share - stock	6241	capital corporation divided individual interest original portion regularly share shares stock
offer - bid	2744	acceptance attempt offer order payment price
stake - share	2661	interest share
business - operation	2643	business functioning mission work
unit - subsidiary	2623	controlled
sale - transaction	2609	

Figure 4.16 Intersection of definitions of similar word pairs from MERGERS

As an example, the list produced for the definition of *administration* shown on page 82 is given in Figure 4.15. Note that in interest of speed no morphological analysis or any other modifications were performed on the tokens in these lists.

When two words are compared in this fashion, the result of the comparison is the intersection of these lists for each word's definition. For example, the intersection between the lists derived from the dictionary entries of *diamond* and *ruby* is (*precious, stone*); between *right* and *freedom* it is (*acting, condition, political, power, privilege, right*).

We performed the following experiment on the same pairs of words derived from the three corpora HARVARD, SPORTS, and MERGERS, as well as on the randomly paired words in RANDOM.

Procedure 3. Given a corpus, take the similarity pairs derived by SEXTANT in order of decreasing frequency of the first term. Perform the intersection of their respective two dictionary definitions as described above. If this intersection contains two or more elements, count this as a hit.

Sample results of intersecting dictionary lists for words judged as similar in the MERGERS corpus are given in Figure 4.16. In this sample the pairs *share-stock*, *offer-bid*, *stake-share*, and *business-operation* are considered hits, since two or more words appear in their intersection, while the pairs *company-concern*, *unit-subsidiary*, and *sale-transaction* are not hits. These examples show the limitations of this evaluation technique, limitations which are inherent in any technique that attempts to mine existing human-directed semantic sources. These sources rely on rich mental representations in the human reader and do not need to include the detail often needed for machine recognition of semantic relations.

Global results are presented in Figure 4.17. They show, as did the evaluation tests with thesauri, that words occurring more frequently in the corpus are more likely to get hits than rarer ones. The reason for this success is similar to that seen in Section 4.3.4, that is, the more frequent words possess better contexts by which to judge similarity. In addition, the more frequent words in the corpus are more likely to be frequent words in general English, and that general words in dictionaries tend to have longer dictionary entries. For example, the words in the first 20 ranking pairs in the SPORTS domain had an average of 82 words in their definition lists, while those appearing in the pairs ranked 1000 to 1020 only had an average of 38 words. The most important result to be seen in the Figure 4.17 for the purposes of this section is that the overlap of the SEXTANT pairs is much greater than the overlap of the pairs formed by randomly chosen corpus words with high frequency corpus words, as in RANDOM, since this shows that **SEXTANT**'s similarity recognition method does indeed provide an overlap with the manually created Webster's that is much better than chance.

4.5 SYNTACTIC VS. DOCUMENT CO-OCCURRENCE

We have already shown during this evaluation against gold standards that using SEXTANT's selective natural language processing to extract context and making comparisons using weighted Jaccard similarity measures permits us to extract semantically related word pairs from raw text. One question that needs to be answered during this evaluation of SEXTANT's performance is what is gained by doing the partial syntactic analysis described in section 3.2. If the same results can be achieved by the more classical methods of examining local context of strings, the reasoning goes, then surely this would be easier to implement and more efficient. In order to demonstrate and measure the

Percentage of hits in Webster's 7th

rank	HARVARD	SPORTS	MERGERS	RANDOM (SPORTS)
1-20	45% (9)	95%	60%	20%
21-40	55% (11)	75%	35%	25%
41-60	70% (14)	65%	50%	20%
61-80	70% (14)	50%	45%	20%
81-100	55% (11)	50%	45%	25%
101-200	55% (55)	53%	31%	15%
201-300	35% (35)	57%	29%	19%
301-400	29% (29)	34%	26%	13%
401-500	25% (25)	48%	22%	14%
501-600	15% (15)	29%	16%	16%
601-700	14% (14)	35%	16%	10%
701-800	16% (16)	34%	15%	11%
801-900	15% (15)	29%	16%	12%
901-1000	11% (11)	21%	11%	16%
1001-2000	6.6% (66)	12.5%	7.4%	9.7%
2001-3000	6.2% (62)	6.6%	3.7%	6.2%
3001-4000	5.2% (52)	5%	3.7%	5.7%
4001-5000	2.9% (29)	5.1%	3.1%	3.6%
5001-6000	2.8% (21)	4.7%	3.3%	2.7%
6001-7000		5.4%		2.4%
7001-8000		4.6%		2%
8001-9000		2.7%		1.8%
9001-10000		2.9%		1.4%

Figure 4.17 Table of hits in *Webster's 7th*

gain that simple syntactic analysis achieves, we decided to create a syntax-free local context baseline in the following manner.

Over a corpus, we perform all the steps normally performed in SEXTANT up to the syntactic disambiguation. In other words, the corpus is divided into lexical units, and each lexical unit is assigned a list of context-free syntactic categories and a normalized form. Everything up until this step is performed by regular grammar analyzers (word division, proper name recognition), or through look-up in a lexicon (morphological analysis). Actually morphological analysis is more complicated than simple look-up, since transformations are performed on word endings; nevertheless, we allow such morphologically analyzed look-up for our base-line system. The next step to be taken by SEXTANT would be to use a disambiguator, which uses the results of a syntactic analysis over the BROWN corpus. Since we are attempting to isolate the effect of syntactic analysis for the baseline here, we use a simpler mechanism, simply considering every word that *can* be a noun as a noun.

SEXTANT uses as the context of a word all the nouns and adjectives that modify that word, as well as the verbs entering into relation with that word. If no syntactic analysis is to be performed, the closest approximation to such connections is to consider any other word within a certain distance of the given word as part of its context. In Gale *et al.* (1992), words within up to fifty words on either side of a given word were found useful in disambiguating word senses, given aligned bilingual corpora. Such a large context is overwhelming for a system whose computational bottleneck is the square of the number of contexts, and we chose to use more modest contexts of 10 words before and after each noun's appearance as part of its context. This is still double the context considered by Phillips (1985). To render the calculation even more tractable, we only compared nouns appearing 10 times or more throughout the corpus. The steps undertaken to produce the baseline non-syntactic similarity measures are:

1. Divide the corpus into lexical units,
2. For each lexical unit look up its possible categories and normalization,
3. Retain all those words which can be nouns and which appear 10 times or more in the corpus,
4. For each noun, retain as its context all the nouns, adjectives, and verbs appearing in the same sentence, not appearing in SEXTANT's stoplist, and within a window of 10 nouns, adjectives, and verbs before or after the word in question,

When Zeiss died in 1888, Abbe took over the firm and established the Carl Zeiss Foundation for scientific research and social improvement; in 1896 he reorganized the entire establishment.

research zeis	research die
research abbe	research take
research firm	research establish
research carl-zeiss-foundation	research scientific
research research	research social
research improvement	research reorganize
research entire	research establishment

Figure 4.18 10 word window context for *research* in a sample sentence

5. Run this data through SEXTANT's similarity module to produce the baseline.

Figure 4.18 shows a sample sentence and the context extracted for one of its nouns.

When we produce this window data for nouns in the HARVARD corpus, 2661 nouns appearing 10 times or more are compared, and 33,283 unique attributes with which to judge the words are extracted. The similarity judging run takes 4 full days on a DEC 5820, compared to 3 and $1/2$ hours for the normal syntactic SEXTANT run, a time increase due to the greatly increased number of attributes for each word.

When we compare the words found closest using this baseline windowing method to those found by SEXTANT's syntactic based method, using the evaluation techniques explained in the previous section, we get the results shown in Figures 4.19 to 4.23. The first table, in Figure 4.19, compares the hits over *Roget's*, the *Macquarie*, and *Webster's*, obtained from the windowing technique described in preceding paragraphs to those obtained from SEXTANT, retaining only words for which similarity judgements were made by both techniques.

It can be seen that the simple technique of moving a window over a large corpus, counting co-occurrences of words, and eliminating empty words, provides a good hit ratio for frequently appearing words since about 1 out of 5 of the 100 most frequent words are found similar to words appearing in the same heading in a hand-built thesaurus. It can also be seen that the performance of SEXTANT is much better for the 600 most frequently

results over HARVARD of Window vs Syntactic Contexts

	Roget		Macquarie		Webster	
rank	window	sextant	wind	sext	wind	sext
1-20	(5 of 20) 25%	50%	15%	40%	55%	50%
21-40	(2 of 20) 10%	30%	20%	45%	40%	60%
41-60	(5 of 20) 25%	30%	30%	35%	55%	70%
61-80	(3 of 20) 15%	30%	20%	30%	45%	65%
81-100	(3 of 20) 15%	40%	15%	35%	35%	55%
101-200	(14 of 100) 14%	31%	19%	34%	34%	55%
201-300	(21 of 100) 21%	29%	20%	30%	29%	34%
301-400	(13 of 100) 13%	17%	12%	18%	25%	29%
401-500	(15 of 100) 15%	16%	12%	13%	24%	26%
501-600	(13 of 100) 13%	11%	10%	15%	19%	16%
601-700	(8 of 100) 8%	11%	11%	14%	20%	14%
701-800	(11 of 100) 11%	9%	9%	9%	17%	17%
801-900	(17 of 100) 17%	6%	13%	7%	25%	12%
901-1000	(8 of 100) 8%	10%	9%	9%	29%	12%
1001-2000	(102 of 1000) 10%	4.9%	11.8%	5.3%	19.2%	6.9%
2001-3000	(33 of 420) 8%	2.4%	7.9%	2.1%	15.2%	5.2%

Figure 4.19 Windowing vs SEXTANT Percentage of Hits for words from most frequent to least

Roget First 600	SEXTANT HITS	SEXTANT MISS
WINDOW HITS	48	60
WINDOW MISS	91	401

$\chi^2 = 6.4$
$p < .025$

Macquarie First 600	SEXTANT HITS	SEXTANT MISS
WINDOW HITS	42	54
WINDOW MISS	103	401

$\chi^2 = 15.3$
$p < .005$

Roget Last 600	SEXTANT HITS	SEXTANT MISS
WINDOW HITS	2	28
WINDOW MISS	14	556

$\chi^2 = 4.6$
$p < .05$

Macquarie Last 600	SEXTANT HITS	SEXTANT MISS
WINDOW HITS	4	40
WINDOW MISS	14	542

$\chi^2 = 12.5$
$p < .0005$

Figure 4.20 χ^2 results comparing SEXTANT and windowing hits in man-made thesauri

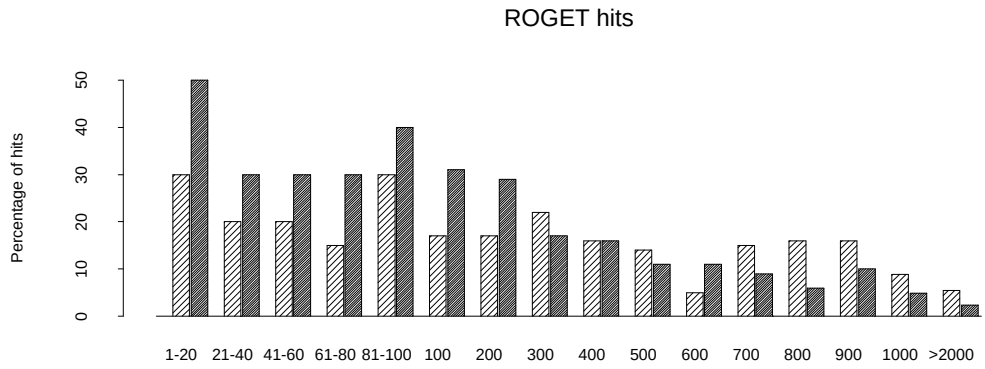

Figure 4.21 Hits in *Roget's*. Comparison of hit percentage in *Roget's* using simple 10-word windowing technique (hashed bars) and SEXTANT's syntactic technique (solid bars). The y-axis gives the percentage of hits for each group of frequency-ranked terms. The solid bars show SEXTANT results, and the hashed bars show the results using windows.

appearing words. The difference in performance between the two techniques is statistically significant ($p < 0.05$). The results of a χ^2 test are given in Figure 4.20. Figures 4.21 to 4.23 show the same results as histograms. In these histograms it becomes more evident that the window co-occurrence techniques give more hits for less frequently occurring words, after the 600th most frequent word. One reason for this can be seen by examining the 900th most frequent word, *employment*. Since the windowing technique extracts up to 20 non-stopwords from either side, there are still 537 context words attached to this word, while SEXTANT, which examines finer-grained contexts, only provides 32 attributes.

This dichotomous results suggests that no one statistical technique is adapted to all ranges of frequencies of words from a corpus. Everyday experience suggests that frequently-occurring events can be more finely analyzed than rarer ones. In the domain of empirical linguistics, the same reasoning can be

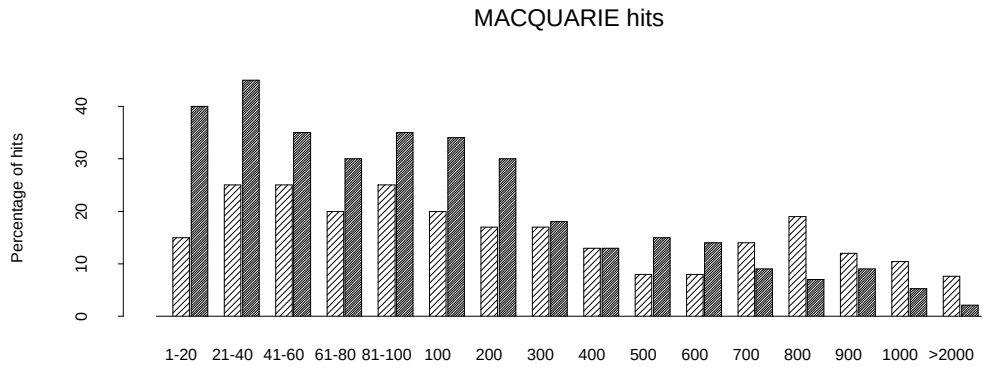

Figure 4.22 Hits in *Macquarie's*. Comparison of hit percentage in *Macquarie's* using simple 10-word windowing technique (hashed bars) and SEXTANT's syntactic technique (solid). The y-axis gives the percentage of hits for each group of frequency-ranked terms.

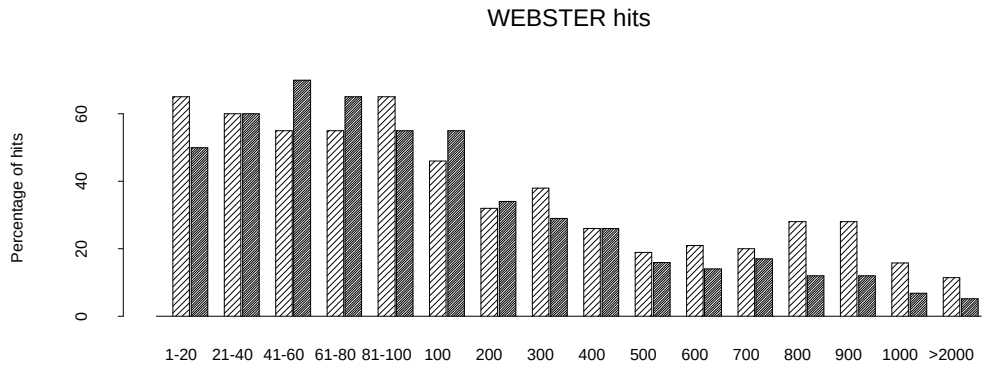

Figure 4.23 Hits in *Webster's*. Comparison of hit percentage in *Webster's* using simple 10-word windowing technique (hashed bars) and SEXTANT's syntactic technique (solid bars). The y-axis gives the percentage of hits for each group of frequency-ranked terms.

Similar pair		*freq of 1st*	*freq of 2nd*
plasma	blood	3	43
recital	voice	3	26
sacristy	nave	3	32
seacoast	coast	3	100
sheep	cattle	3	16
slavic	baltic	3	7
subgroup	tribe	3	63
tapestry	piece	3	57
beach	bank	4	1462
bubble	flow	4	37
circuit	province	4	107
corridor	location	4	50
duration	period	4	350
initiate	member	4	588
jurisprudence	legislation	4	92
lyceum	academy	4	478
mantle	crust	4	20

Figure 4.24 Frequency of pairs of similar words discovered by SEXTANT and matching a thesaurus. Some low frequency words can be accurately matched to higher frequency words.

applied. For frequent words, finer-grained context such as that provided by SEXTANT, is rich enough to judge similarity. For less frequent words, reaping more but inexact information, such as that given by windows of N words, provides more information about each word. For rare words, the context may have to be extended beyond a window, to the paragraph, or section, or entire document level, as Crouch (1990) did for rarely appearing words.

It may be tempting to divide the words of a corpus into groups according to frequency, and to apply different techniques to each group. But an examination of the similar pairs that were discovered by SEXTANT and matched against *Roget's* or *Macquarie's* show that words can be accurately matched across frequency groups, as can be seen in the sample from the HARVARD corpus given in Figure 4.24. This sample shows low-frequency words which scored hits with words from a variety of frequency groups. Since such hits are still possible at low frequencies, and since results for low-frequency words in Figures 4.14 and 4.17 are better than random pairings, SEXTANT findings for low-frequency words need not be dismissed out of hand but that, rather, independent confirmation via results from other corpora

or from other knowledge-poor techniques may allow these relations to be confirmed and thus reclaimed.

4.6 SUMMARY

In this chapter we have examined three distinct methods for confirming that the relations extracted by SEXTANT are semantic relations between words. Common adjectives frequently associated by human subjects are often associated by the mechanical, knowledge-poor techniques of SEXTANT. Artificially created synonyms can be found by SEXTANT. And, finally, from raw text SEXTANT extracts word-pairs that are often found under the same subheading of man-made thesauri. In each of these methods, results from SEXTANT were better for the characteristic vocabulary of the corpora tested, i.e., those words which appear throughout the corpus and which possessing a large number of clues as to their meaning. Results in this chapter also suggest that different approaches to knowledge-poor semantic extraction should be conjugated. For rare words, for example, techniques using contexts such as large textual windows give better overlap with existing thesauri. The overall conclusion of this chapter is a confirmation that knowledge-poor techniques such as the selective natural language processing of SEXTANT can extract semantically similar words from raw text.

5

APPLICATIONS

In the last two chapters, we described our semantic extraction techniques and showed that the list of similar words extracted corresponded to the types of lists manually created by humans for general English. We argued that the advantage of having an automatic technique that approximates such extraction is that, in addition to being fast and economical, it provides information that is specific to the corpus from which it is derived. Here we present a few possible applications of these extracted relations. In the next section, we describe our experiments with the automatic expansion of queries in a classical information retrieval setting. After that, we present experiments showing how the techniques developed in SEXTANT can be applied to enriching existing knowledge structures. We treat the problem of inserting a new word into its proper place in a thesaurus. These experiments also demonstrate how two knowledge-poor techniques can reinforce each other. Then we show how a deeper exploration of the information extracted by SEXTANT permits the creation of clusters of words along semantic axes. Finally in the last section of this chapter, we organize all the disparate techniques developed throughout this book and demonstrate how the first draft of a corpus-derived thesaurus can be automatically created from raw text.

5.1 QUERY EXPANSION

A classical information retrieval system follows the following paradigm. The data base is composed of free text units called documents. A query is a natural language expression of interest that may consist of a few key words or phrases, or maybe a number of sentences, describing what the human user would like to

retrieve from the data base. The information retrieval system digests the query and produces a ranked list of documents responding to it. The most common implementation of such a system consists of a list of keywords indexing the documents and a mechanism that matches keywords extracted from the query with the keywords in the index. Common matching strategies often employ stopwords lists that eliminate insignificant words, and then they match the remaining words, or truncated versions of the remaining words, from the query list and document lists. The matching can entail a simple count or can weight the words, using their frequency in the query and documents. The mechanism can then calculate a closeness measure between the weighted query words and the documents.

A common similarity measure is the cosine measure, in which each word (or truncated word) from the language is considered to form an axis in a large dimensional space. The presence of a word in a document or query signifies a magnitude in the direction of that axis. Any query or document can be considered as a vector in this space, whose position is determined by its magnitude (determined usually by the frequency of the word) along each axis. The cosine measure between a query and a document is the cosine of the angle between their two vectors, which is equal to 1 if they are in exactly the same direction, and zero if they are orthogonal.

This geometrically based theory is very attractive and has been shown to be better than simple boolean matching of keywords. Indeed, almost 25 years of research exploring this and other knowledge-poor models has resulted in little improvement over this simple scheme (Salton & McGill 1983).

One theoretical drawback to the cosine model is that it presupposes that the axes are orthogonal and independent. When keywords were assigned by hand, as was the case in early information retrieval systems, there was the hope that orthogonality could be maintained by the human indexer, although this seems not to be the case (Sievert & Andrews 1991) in large scale hand indexing. Axes are certainly not orthogonal for automatically extracted keywords, even stemmed words, as the phenomenon of *language variability*, discussed in Section 1 demonstrates.

Some responses to this drawback have been to try to reduce the dimensionality of this space from $O(N)$, where N is the number of words in the language, to a smaller number, using singular value decomposition (Deerwester *et al.* 1990; Dumais 1993), using predefined semantic codes (Liddy & Paik 1992), using discovered equivalence classes (Salton 1972), or using even simple stemming (Porter 1980).

name	*Documents*	*size (KB)*	*queries*
ADI	82	39	35
CACM	3204	1300	64
CISI	1460	1300	112
CRAN	1400	1400	225
MED	1033	1000	30
NPL	11429	3200	100
TIME	425	1500	83

Figure 5.1 Traditional Information Retrieval Testbeds

Another response is to expand the initial query with words related to the original query word in the corpus (Sparck Jones & Barber 1971; Minker *et al.* 1972; Harman 1988), thus making the original query vector closer to more documents. This is of course the dual of dimension reduction, since expansion of words can be seen as collapsing of axes in the word space. Since we claim that SEXTANT is able to find similar words by examining the syntactic contexts of words in a corpus, it is natural to try using these similar words as expansion candidates for queries on that corpus.

In this section, we report on the results of query expansion using SEXTANT derived word lists. We find that performance for some queries is improved, while for other queries precisions goes down. By examining both sets we try to explain in what contexts query expansion may be useful.

5.1.1 Testbeds

A number of limited testbeds were developed over the ’60s and ’70s for evaluating information retrieval systems from this system-oriented perspective, and these are available in the public domain. Most of the research cited in the previous paragraphs has been performed on these corpora[1] listed in Figure 5.1.

There are severe limitations involved in using these corpora.

1. Small size corpora: Most are about 1 megabyte, or about 200,000 words.
2. Articles all in one case: Some corpora, such as TIME are all upper case, frustrating case-based methods of recognition of name units, which then perturb subsequent syntactic analysis.

[1] See Appendix for more detailed description of these corpora.

3. Computer illiterate queries: For example, ‘‘How can machine translating compete with traditional methods of translating in comprehending nuances of meaning in languages of different structures?’’ Such questions presuppose high level reasoning and language understanding of which no computer today is capable.
4. Short documents: For example, NPL ‘‘documents’’ are merely titles sometimes only three words long.
5. Diverse subject areas: For example, the MED database contains abstracts from psychology, cardiology, pediatrics, etc.

Despite the flaws in these widely used testbeds, we feel obligated to demonstrate our technique on them as a comparison point to former research. Incidentally, a series of new and larger testbeds are currently being prepared through the series of TREC (Harman 1993) conferences organized by the National Institute of Standards and Technology.

The traditional approach to automatic discovery of word relatedness (Salton 1971) has been to use an entire document, e.g., an abstract, as the context of a word which appears in it; such a context is easy to extract automatically. The document co-occurrence hypothesis is that two words appearing in the same document share some semantic relatedness. A number of papers have called into doubt the usefulness of document-co-occurrence-derived similarity (Minker *et al.* 1972; Sparck Jones 1991; Peat & Willet 1991). Problems associated with document co-occurrence have been discussed in Section 2.4.3 (see page 27).

5.1.2 Traditional Evaluation Techniques in IR

Each of the traditional testbed corpora possesses its own manually created list of queries and relevant documents that answer them. The classic evaluation methods over these corpora involve treating and storing the documents, automatically treating each question, and matching the treated queries with the treated documents. This matching produces an ordered list of documents considered relevant to each query.

Using the human-built relevance list supplied with the testbed corpus, it is possible to go down the ranked list produced by the system, marking which document retrieved is actually relevant or not. For each query, a measure of recall (the percentage of relevant documents found), as well as precision (the

percentage of documents recalled that were relevant) can thus be measured. This measurement is usually done at different levels of recall. For example, suppose that 30 documents have been manually judged relevant to a given query in one of these testbeds. Now suppose that some automatic retrieval system processes this query and returns a ranked list of documents for it. If the first five documents contained three relevant documents and two irrelevant documents with the fifth document being one of the relevant documents, then at a 10% recall level (that is, 3 of the 30 documents recalled) we would a 60% precision measurement (since 3 of the 5 documents recalled were relevant). If the first 40 documents returned contained 27 relevant documents, then at the 90% recall level we would have a $27/40$ or 67.5% precision measurement. These result of an automated information retrieval system are often given as plots of precision at certain recall levels, e.g., 10%, 20%, 30%, ..., 90%; sometimes an average of these nine levels or of three intermediate levels (25%, 50%, 75%) is given as an overall measure of precision.

5.1.3 Our experiments

We ran query expansion experiments over all of the testbeds mentioned in Figure 5.1. Our baseline test consisted of producing a representation of the documents by eliminating stopwords and morphologically normalizing all the remaining words. The same treatment was applied to each query. Then each query was converted into a weighted vector, using the inverse log of the frequency of the word in corpus to weight each query word.

$$weight(word) = \frac{1}{log_2(freq_{corpus}(word) + 1)} \tag{5.1}$$

This is a classic weighting technique that has been shown to improve the overall precision results (Salton 1971). Then a cosine measure of the resulting query vector was calculated against all the weighted document vectors, using the programs that we developed for Chapter 3; next the documents were ranked from closest to furthest. Documents sharing no terms with the query were not included in the ranking. This gave us a baseline result for each corpus. The results we obtained were similar to those published in (Salton 1971).

As variations on this baseline, we augmented the queries using (1) discovered families of words which are discussed below, (2) words found to be close using a similarity calculation based on document co-occurrence, and (3) words found to be close based on SEXTANT analysis of the corpus. In addition, runs were performed on stemmed versions of the baseline and of the expanded

queries. The stemming algorithm is described in Porter (1980). Results for each run and combination were evaluated using the classic nine precision level technique described above, as well as with a different technique described in (Croft 1993) which measures the precision at fixed numbers of documents.

When the queries were expanded, only the closest group of words was used to expand a query term. The closeness of the words was calculated by using the Jaccard measure which ranks words from 0 (closest) to 1 (furthest) as described in Section 3.2.6. We define our 'closest' groups as the most similar (closest) word as well as any word within a distance of 0.01 of that word.

5.1.4 Word Families

Before giving the results of our query expansion experiments we first describe a simple process that finds families of words using corpus context and a string matching procedure. During our experimentation, we found that using as the context of each word the document numbers that it appears in often groups morphological variants of words together as being close. The same phenomenon appears when sentence numbers are used as context, but to a lesser degree. At first we were tempted to consider this as noise, until we realized that these variants may be useful in query expansion, as experimentation below shows.

That variant forms of the same word should appear in the same document is normal, since the same concept may be expressed as a noun, a modifier, or a verb. Such uses imply morphological transformations in English and most natural languages. These morphological variations are numerous and often domain dependent; e.g., medical terminology follows different formation rules than finance. If the morphological rules of the domain have not been sufficiently analyzed and codified, it would be interesting to have a procedure for discovering them automatically. It seems that document co-occurrence, Jaccard similarity, and a matching heuristic provide the elements necessary to perform this discovery.

We developed a family discovery program using the document number that word appears in as its context and the heuristic matching algorithm in Figure 5.2 to extract family variants. Words are considered close by this algorithm if they appear in the same documents and share the first three, four or five letters, depending on word length. This algorithm favors non-initial morphological variation, although matchings could have been based on more complicated

```
For each V among the closest words to W
    lengthV = stringlength(V)
    lengthW = stringlength(W)
    minLeng = MIN(lengthV, lengthW)
    if (minLeng < 3) CompareLeng = minLeng
    else if ( minLeng <= 5 ) CompareLeng = 3
                else if ( minLeng <= 10 ) CompareLeng = 4
                else CompareLeng = 5
    if V and W agree in their first CompareLeng characters,
             ACCEPT V as being a variant of W
```

Figure 5.2 Algorithm for Selecting Morphological Variants of the same word

pattern matching such as letter bigram matching (Adamson & Boreham 1974). Using this scheme over the MED corpus gives word variants such as those shown in Figure 5.3.

Some of the examples given in the table are erroneous, such as the pairs *acidic-ac*, *actinomycin-act* and *valvulography-valid*. In these three cases, at least one of the words appears in only one document. Other pairs, such as *autograft-autochthonous* or *cerebrospinal-cerebral* may have some semantic closeness, but certainly not the tight relation that might permit them to be classified as belonging to the same family. Nonetheless, we included them in our automatic expansion technique.

5.1.5 Experimental results

Figures 5.4 and 5.5 show the results of expanding the query in the MED testbed[2] using the techniques discussed above. The seven columns show results from seven different querying techniques. The *Base* case is obtained by eliminating stopwords from the queries, morphologically normalizing the resulting words through dictionary look-up[3], and calculating the cosines between the resulting vector and the vectors obtained for each document also obtained by morphological normalization. This allows the ranking of documents against each query, a ranking whose precision is verified against

[2] For the other testbeds see the description of the corpus in Appendix 6.

[3] Note that this morphological treatment which reduces plural nouns to singular nouns, comparative adjectives to simple adjectives, and conjugated verbs to infinitives is slightly more sophisticated than the standard IR paradigm which does not perform this normalization.

abnormality abnormal	...
acetoacetate acetate	sth-maintained sth
acidic ac	stone-forming stone
acromegaly acromegalic	strainspecific strain-specific
actinomycin act	streptococcus streptococcal
adeaminase a-deaminase	suicide suicidal
adhesiveness adhesion	sulfokinase sulfate
adrenergic adrenalectomized	sulfur-containing sulfur
alaly alalies	surfactant surface
amyloidosis amyloid	surgical surgery
aneurysmal aneurysm	symbiotic symbiosis
angiographic angiocardiography	tenuis tenuazonic
antecendent antecedent	teratogenicity teratogenic
anticancer anti-tumor	therapy therapeutic
antigenic antigen	thrombin thrombase
antimetabolite antidote	thrombocytopenia thrombin-fibrinogen
antisera antigenic	thrombotic thrombosis
antiserum anti-hgh	thymectomy thymectomized
antiserum antibody	thymidine-h thymidilate
atmosphere atm	thymo-lymphatic thymic
atrioventricular atria	thymocyte thymic
atrium atria	thymus thymectomized
autistic autism	thyroidectomized thyroid
autograft autochthonous	thyroxin-stimulated thyroxin
autolytic autolysis	toxemic toxemia
autopsy auto-immunization	toxicity toxic
bacteriophage bacterial	toxicosis toxic
bacterium bacterial	transportation transference
bacterium bacteriophage	transposition transportation
biliary bile	triphosphopyridine triphosphatase
breakage break	triphosphopyridine triphosphate-inorganic
british britain	triturus triton
bronchogenic bronchial	tubule tubular
bronchogenic bronchioloalveolar	tumoral tumor-like
bu-15 bu-100	urea-treated urea
c.s.f. c.s.	uremic uremia
calciphylaxis calciphylactic	ureteral ureter
calvarial calvaria	uvr uv
cancerous cancer-specific	valvular valve
carcinogenic carcinogen	valvulography valid
carcinogenic carcinogenesis	veno-arterial vena
cataractous cataract	ventricular ventricle
cation ca	ventriculoatrial ventriculo-peritoneal
cerebrospinal cerebral	ventriculovenous ventriculo-venous
cerebrovascular cerebro-vascular	virus viral
chelate-enzyme chelate	virus-specific virus-infected
chelation chelate	x-ray-irradiated x-ray
childhood child	xanthoma xanthogranuloma
chlorothiazide-induced chlorothiazide	xirradiated x-irradiated

Figure 5.3 Sample Word Variant Discovered in the MED corpus Using Document Co-occurrence

the human-created lists of relevant documents per query that are supplied with the testbed.

The other six techniques modify the query vector. The technique marked DOC augments query terms with terms that were found closest in the document set, using document co-occurrence as context for judging similarity (Harman 1988). The SEXT column shows the results of augmenting query terms with terms found most similar using SEXTANT. STEM involves stemming the original query using the Porter algorithm (Frakes 1992b), and calculating cosines against vectors from a stemmed document base. FAM shows the results of augmenting the query with its family terms, derived as described on page 106. S+FAM augments the queries with SEXTANT-discovered similar words, and then family words are added to the queries. The S+F+STEM technique does the same, and then stems the resulting augmented query.

For each technique, the average precision over the 30 queries is given at recall levels of 10%, 20% to 90%. The average of these nine recall levels is also given. For example, the average precision of the SEXT method when half of the relevant documents had been found in each query was 57.5%. The average precision over all nine recall levels for this technique is 53.4%.

These results show that the average precision of the queries improves when the queries are augmented by the closest words found similar by SEXTANT. The average precision goes from 0.511 to 0.534, and this improvement is greater than that derived from adding in words found closest by using document co-occurrence as context (0.524). This measure of improvement in average precision over many queries (30 here) has been the classical measure (Frakes 1992a) of performance of information retrieval systems (see (Harman 1992; Keen 1992) for some recent examples). The best results, 54.0%, in this category over the MED testbed are attained by automatically augmenting queries with the terms found to be closely similar by SEXTANT, then adding in all words found to be in the family of the augmented query terms, then stemming the resulting query, and applying it to a stemmed version of the document set. The gain from all this work, although done all automatically, is a modest 5-6%. Results for the other traditional testbeds are shown if Figure 5.6. This table shows the number of queries[4] whose performances improve or deteriorate after augmentation by family members and by SEXTANT-discovered similar words on a stemmed testbed. Performance results on these testbeds are also mixed.

[4] The best and the worst queries appear after each testbed corpus description in the Appendix.

	MED						
	Base	Doc	Sext	Stem	Fam	S+Fam	S+F+Stem
	PRECISION						
Recall: 10	0.717	0.806	0.718	0.733	0.704	0.711	0.737
Recall: 20	0.671	0.703	0.696	0.666	0.660	0.664	0.654
Recall: 30	0.641	0.649	0.644	0.645	0.628	0.621	0.639
Recall: 40	0.608	0.612	0.623	0.604	0.580	0.609	0.610
Recall: 50	0.537	0.553	0.575	0.535	0.550	0.561	0.570
Recall: 60	0.484	0.457	0.536	0.477	0.497	0.514	0.517
Recall: 70	0.407	0.393	0.447	0.403	0.441	0.474	0.478
Recall: 80	0.332	0.334	0.351	0.339	0.349	0.399	0.394
Recall: 90	0.202	0.211	0.216	0.200	0.226	0.263	0.261
Average	0.511	0.524	0.534	0.511	0.515	0.535	0.540
Better	---	14	13	16	11	14	17
Same	---	0	9	2	7	5	2
Worse	---	16	8	12	12	11	11
	RECALL						
At 5 docs:	0.53	0.59	0.54	0.54	0.50	0.51	0.51
At 10 docs:	0.56	0.56	0.57	0.55	0.56	0.57	0.56
At 15 docs:	0.54	0.53	0.56	0.54	0.52	0.55	0.56
At 20 docs:	0.50	0.50	0.51	0.52	0.50	0.51	0.51
At 25 docs:	0.47	0.46	0.48	0.47	0.47	0.47	0.47
Better at 15	---	10	9	7	6	9	12
Same at 15	---	7	18	18	12	15	11
Worse at 15	---	13	3	5	12	11	7

Figure 5.4 Query Expansion Results over the MED Testbed.

MED --- BEST IMPROVEMENTS

Base Query	*Augmented Query*	*change*
infantile autism	infantile **infancy** autism **psychosis schizophrenia autistic psychotic schizophrenic schizophrenics**	0.399 to 0.795
homonymous hemianopsia visual aphasia measurement assessment gerstmann syndrome agnosia	homonymous hemianopsia **hemianopia** visual aphasia measurement assessment gerstmann syndrome agnosia **agnosic**	0.476 to 0.706
palliation temporary improvement cancer patient drug x-ray surgery	palliation **palliative** temporary improvement cancer **carcinoma** patient **case** drug x-ray surgery **operation operative surgical**	0.549 to 0.727
renal amyloidosis complication tuberculosis effect steroid condition term kidney disease nephrotic syndrome select requester prednisone prednisolone steroid	renal amyloidosis **amyloid** complication tuberculosis **tb** effect steroid condition term kidney disease nephrotic **nephotic** syndrome select requester prednisone prednisolone steroid	0.609 to 0.727

MED --- WORST RESULTS

Base Query	*Augmented Query*	*change*
nickel nutrition requirement analysis enzyme system toxicity human laboratory animal deficiency sign symptom level foodstuff level blood tissue	nickel nutrition requirement analysis enzyme system toxicity **toxic** human laboratory animal **mice** deficiency **deficient** sign symptom level **concentration** foodstuff level **concentration** blood tissue	0.203 to 0.129
induce hypothermia heart surgery neurosurgery head injury infectious disease	induce hypothermia **hypothermic** heart surgery **operation operative surgical** neurosurgery head **lymphoid lymph-node lymphocyte** injury infectious disease	0.307 to 0.232
ventricular septal defect occur association aortic regurgitation	ventricular **ventricle** septal **septum** defect occur association aortic **aorta** regurgitation **insufficiency regurgitant**	0.808 to 0.700
blood urinary steroid human breast prostatic neoplasm	blood urinary **urine** steroid human breast **adenocarcinoma** prostatic **prostatectomy** neoplasm	0.747 to 0.630

Figure 5.5 Best and Worst Results from Query Expansion on MED.

name	*Documents*	*size (KB)*	*queries*	Better / Same / Worse	Avg. Improv.	Theor. Improv.
ADI	82	39	35	16 / 7 / 12	10.4%	12.2%
CACM	3204	1300	64	17 / 3 / 32	-17.7%	4.7%
CISI	1460	1300	112	31 / 3 / 42	0.0%	4.5%
CRAN	1400	1400	225	59 / 11 / 155	-5.6%	6.8%
MED	1033	1000	30	17 / 2 / 11	5.7%	9.3%
NPL	11429	3200	100	34 / 3 / 56	-12.5%	3.1%
TIME	425	1500	83	30 / 27 / 26	0.0%	13.8%

Figure 5.6 Results of Similarity-Based Expansion on Traditional Information Retrieval Testbeds. For each testbed, the table gives the number of queries whose performances improve, stay the same, or deteriorate after augmentation with family and similar words on a stemmed database.

The reasons for these mixed results, where the expansion improves some query results and degrades others, seem to be the following:

1. Indiscrimination between modifiers and head nouns in queries. In the treatment of query terms each individual word is expanded as if it were the only word in the query. An example of this, which may be the reason for poor performance, is given in one of the questions displayed in Figure 5.5, in which the word *head* is expanded by *lymphoid*. Whereas it is conceivable that someone querying about the head would be interested in the lymph glands that are located under the jaws, in this query the thrust of the query concerns *head injuries*, and more specifically *hypothermia* after such injuries, and the word *head* enters in as the location of interest. Here *head* plays a restrictive modifying role yet it is expanded as if it were a head noun of the phrase. If automatic expansion is to be attempted, perhaps different expansions should be given for words appearing in a limiting, modifying capacity. Of course, the larger problem is recognizing the general *topic* of any discourse (Wilensky 1992).
2. Indiscrimination along semantic axes. As seen in Chapter 4, the relations extracted as similar by SEXTANT can be antonymous relations between words at different poles along the same semantic axis. For example, the word closest to *female* in most corpora we treated was *male*. But when a person uses the word *female* in a query it is usually in opposition to *male* and thus it would be improper and counterproductive to expand *female* automatically by *male*.

The solution to both these problems, which boils down to knowing what the interrogator is really looking for, would probably be to consider query expansion in an interactive mode. A user interrogating a textual corpus might like to browse among a list of similar words as possible expansions. This technique has been suggested before (Nelson 1993), but with the browsing lists drawn from some man-made source such as thesauri or dictionaries, and not from relations found within the text itself. A visual graphic interface such as that developed for closeness between documents in Korfhage (1991) might well be adapted to displaying the relationship between terms themselves in the corpus.

Although it would seem that the techniques developed by SEXTANT for recognizing similarity between words provide only mixed results when applied in a brute force manner to query expansion, additional human interaction may allow an information retrieval system to profit from these techniques. To

give an idea of what might be gained by implementing this round of human interaction, we present in Figure 5.6 the results of applying the augmentation techniques in an ideal setting where a human chooses only those expansions that improve the query's performance. The column marked THEOR. IMPROV. shows how much improvement may be gained by retaining only those queries whose performance improves after automatic expansion. These theoretical improvements range from a 3% improvement to 13% improvement in the average precision of the documents recalled per query.

5.2 THESAURUS ENRICHMENT

Interest[5] in knowledge-poor techniques for extracting information from text is growing (Zernik 1991; Weir 1992; Goldman 1992; Boguraev 1993). Wilks *et al.* (1992) discuss the potential power behind combining weak methods and describe advances achieved using this paradigm. In this section, we show how SEXTANT may be combined with a very different technique to enrich a manually created thesaurus[6] for . Although the results are modest, they indicate again how different knowledge-poor techniques may be overlaid to obtain stronger results than either one separately is capable of producing.

Robison (1970) first suggested that lexico-syntactic patterns may provide a rich ground for extracting automatic semantic relations. He extracted forty thousand such patterns for English, but his suggestions were not subsequently pursued. As mentioned, Hearst (1992) recently presented an automatic lexical discovery technique that uses lexico-syntactic patterns to find instances of the hyponymy (i.e., ISA) relation in large text bases. For example, consider the lexico-syntactic pattern, where *NP* means a noun phrase and where { }* means that the enclosed may repeat any number of times,

... NP {, NP} {,} or other NP ...*

This pattern is one of several that have been identified as indicating the hyponymy relation. When a sentence containing this pattern is found (with some restrictions on the syntax to the left and the right of the pattern) it can be inferred that the NPs on the left of the phrase ''or other'' are hyponyms of the

[5] This section describes work done with Marti Hearst. An earlier version of this section appeared as (Grefenstette & Hearst 1992).

[6] See Yarowsky (1992) for the inverse problem of using thesaurus headings to disambiguate meanings of words present in the thesaurus.

NP on the right (where NP indicates a simple noun phrase). From the sentence ''Bruises, wounds, broken bones or other injuries are common,'' we can infer:

hyponym(bruise, injury)
hyponym(wound, injury)
hyponym(broken bone, injury)

One interesting aspect of this knowledge-poor method, in comparison to ours (although both require as their primary resource a large text collection), is that only one occurrence of the pattern need be encountered for the relation to be extracted. Indeed, such a lexical syntactic pattern may be the only clue given for a word or term appearing only once in the corpus.

One use of a technique that extracts these explicit hierarchical relations is the enrichment of an existing knowledge source, for example, a manually created thesaurus. This placement may be problematic, though. Suppose that a hypernym relation is discovered by this method which does not exist in the existing knowledge source. The questions remain: At what level of hyponymy should the newly discovered hyponym be placed? If the hypernym possesses many distinct subtrees of hyponyms, in which subtree does the new term go? It is to answer such questions that SEXTANT might be applied for words that appear with a certain frequency in the corpus that are new hyponyms for the knowledge base.

We decided to perform our experiments uniting these two methods using the manually constructed thesaurus WordNet (Miller *et al.* 1990). Word forms with synonymous meanings have been manually grouped into sets, called `synsets`, in WordNet. These synsets are similar to the `rows` in Spark Jones's early work mentioned on page 18. This presentation allows a human to distinguish between senses of homographs. For example, the noun *board* appears in the synsets {*board, plank*} and {*board, committee*}, and this grouping serves for the most part as the word's definition. In version 1.1 that we used for these experiments, WordNet[7] contained about 34,000 noun word forms, including some compounds and proper nouns, organized into about 26,000 synsets. Noun synsets are organized in a directed acyclic network according to the hyponymy relation with implied inheritance and are further distinguished by values of features such as meronymy. WordNet's coverage of general English is extensive although, as with any thesaurus, it covers only

[7] WordNet is currently (February 1993) in its version 1.3. It is freely available via anonymous ftp from clarity.princeton.edu. The entire package is 12 megabytes.

a small part of possible relations. It certainly provides a good base for an automatic acquisition algorithm to build on.

Taking WordNet as our base then, we assume that we have discovered the relation *hyponym(A, B)*, indicating that *A* is a kind of *B* in some corpus and that we wish to enrich the WordNet network with this relation. If there are subtrees of hyponyms under *A*, then the above questions about placement of *B* must be answered.

It has been observed (Kelly & Stone 1975; Gale *et al.* 1992) that the sense of a word can sometimes be inferred from the lexical contexts in which the word is found. As a simple example, when *bank* is used in its riverbank sense, it is often surrounded by words having to do with bodies of water, while when used in its financial institution sense, it appears with appropriate financial terms. In a similar way, we plan to compare the syntactic contexts of the words in each subtree of the hypernym of interest with the context of the new hyponym, then place it in the subtree to which it is most similar.

5.2.1 Corpus Extraction

As an experiment, we use one of the relations *hyponym(Harvard, institution)* extracted from an on-line encyclopedia (Grolier 1990) using the technique described in Hearst (1992). As noted above, if we wish to insert this relation into a hierarchical structure such as WordNet, we have to decide which sense of *institution* is appropriate. In Figure 5.7, we list an abbreviated version of the WordNet synsets associated with the various senses of *institution*. Each sense is followed by the hyponymic subtrees associated with it, indicated by an arrow.

Our goal is to see if, by examining the syntactic contexts of these terms in a corpus of text, we can decide under which synset to place *Harvard*.

Given a large enough text sample, SEXTANT can tell us what words are used in the most similar ways to *Harvard*. In order to generate this text, we took all the individual words from the above list, giving the list of words:

> institution, establishment, charity, religion, faith, church, vicariate, vicarship, school, educational, academy, honorary society, foundation, bank, commercial bank, orphanage, orphans' asylum, penal institution, constitution, establishment, formation, initiation, founding,

```
Institution: (hyponyms)
institution, establishment
 => charity
 => religion, faith, church
              => vicariate, vicarship
              => school, educational institution
              => academy, honorary society
              => foundation
              => bank, commercial bank
institution
              => orphanage, orphans' asylum
              => penal institution
constitution, establishment, formation, initiation, founding,
foundation, institution, origination, setting up, creation,
instauration
              => colonization, settlement
```

Figure 5.7 WordNet entries under the word *institution*

foundation, institution, origination, setting up, creation, instauration, colonization, settlement.

We extracted all the sentences from Grolier (1990) that contained those terms. This generated the 3.9 MB HARVARD corpus[8].

As described in Chapter 3, we processed this corpus, performing morphological analysis, dictionary look-up, grammatical disambiguation, division into noun and verb phrases, parsing, and extraction of lexical attributes for each noun and adjective in the corpus. In this experiment, we decided to use the context of each word, not only when it was a head noun, but also when it was a modifier, as was done in Section 4.1. The reason for this expansion is that after a first run it was found that *Harvard* appeared rarely as a head noun. This choice is discussed below.

As a sidenote to these experiments, it turns out that the extraction technique using WordNet hyponyms sets provides an interesting slice of the corpus. In the HARVARD corpus, for example, by extracting all phrases containing one the hyponyms of *institution* from a general encyclopedia, it seems that we induce

[8] This corpus is more fully described in Appendix 6.10.

	HARVARD		SPORTS	
word	*freq*	*closest terms*	*freq*	*closest terms*
action	138	function act reform	310	pressure energy behavior
body	230	organization number group	658	area group part
career	175	training study curriculum	392	history record team
case	130	law provision	372	decision issue
child	375	student education people	188	individual people person
claim	125	sovereignty control majority	129	right protection gain
composition	105	fresco music book	183	setting space quality
condition	126	standard need opportunity	267	environment temperature factor
course	154	training curriculum education	185	way plan channel
deposit	157	account reserve resource	241	rock lake oil
difference	104	identity orientation conflict	173	variation change increase
effect	141	impact change contribution	467	change energy pressure
element	220	structure form influence	314	material substance style
event	113	figure aspect issue	355	competition race activity
experience	187	life approach relationship	130	knowledge analysis psychology
facility	164	education service agency	163	opportunity resource dam
field	109	curriculum degree issue	3646	surface water system
freedom	262	right independence status	110	privilege prohibition
function	168	purpose belief authority	285	activity structure need
importance	105	success need emphasis	141	popularity significance attention
land	217	property territory place	396	region resource soil
level	222	rate curriculum education	489	temperature pressure amount
life	452	history tradition society	606	activity population group
line	162	nature control view	531	track number point
member	588	group year constitution	501	man player school
opportunity	129	suffrage skill education	101	impetus facility charter
order	395	community life belief	268	power number way
organization	322	agency institution community	245	education individual school
play	131	poetry poem book	2407	work role game
point	135	piece proponent concern	573	temperature number amount
policy	304	reform position authority	180	need service resource
position	264	post role policy	282	post condition location
practice	382	education tradition belief	222	tradition school application
principle	210	idea law belief	183	rule technique basis
rate	174	level number tax	212	temperature density strength
science	216	education history psychology	166	research knowledge education
system	1292	institution program government	1556	field form area
theory	329	idea view thought	422	study art force
tradition	423	art religion belief	241	art style artist
training	266	education curriculum course	149	education instruction school
view	242	idea interpretation doctrine	162	interest picture diversity
wall	125	chapel roof space	285	layer side portion
war	263	colonization conflict struggle	349	fight force king
work	1028	school institution century	1165	play game study

Figure 5.8 Terms Similar to Terms ‘‘once-removed’’ from *institution*

school	study-SUBJ	school-MOD	university-MOD
law-MOD	faculty-MOD	college-MOD	graduate-MOD
review-MOD	divinity-MOD		

Figure 5.9 Attributes of both *Harvard* and *Yale* in the HARVARD corpus.

an *institutional* sense on the other words extracted. Consider Figure 5.8. There we show the results of performing our similarity analysis on these terms ‘‘once-removed’’ from the list of institutional terms used to generate the corpus. From this table, it appears that the institutional sense of some polysemous words has been identified, e.g., *member-group, right-freedom, service-program,* and *union-association.* In other words, it seems that by taking a WordNet synset as a filter through a large text, we have extracted a rather coherent corpus. In this case, the corpus is one with a bent towards the notion of institution. In a corpus consisting of articles about tennis, we would expect different associations for words like *service, movement,* and *court*.

This observation helps substantiate the claim that we can use existing knowledge structures to help coarse-level analysis techniques. In this case the thesaurus helps find a semantic partition on unrestricted textual data.

5.2.2 ‘Harvard as an Institution’ Experiment

Many of the terms in the institution synsets were found to have reasonable associations. Figure 5.10 shows, for each word listed in WordNet as an immediate hyponym of *institution*, the word whose lexico-syntactic context was most similar to it among all of the 22,000 unique words examined. Terms associated with words with low frequency (such as *orphanage* and *asylum*) tend to be less plausible than higher frequency words.

As for our original concern as to where to place *Harvard* as an *institution*, SEXTANT finds that *Harvard* is used most similarly to *Yale, Cambridge, Columbia, Chicago, Oxford, and Juilliard*. For example, *Harvard* and *Yale* are found to be similar because they both modify or are both modified by the collection of words shown in Figure 5.9.

The fact that SEXTANT places *Harvard* as being most similar to other university names, but not the term *university* itself, points out one difficulty with our task. This particular portion of WordNet does not contain listings of

term	*freq*	*closest term*	*closest* INSTITUTION *hyponym*
establishment	1140	creation	creation
charity	76	devotion	colonization
religion	2347	religious	faith
faith	835	religion	religion
church	8308	school	school
vicariate	3	prothonotary	faith
school	10012	institution	church
academy	2254	university	school
foundation	1697	institution	school
bank	2612	institution	settlement
orphanage	11	yverdon	asylum
asylum	11	promulgation	orphanage
orphan	4	mottel	---
penal	16	roanoke	colonization
constitution	2062	state	church
establishment	1140	creation	creation
formation	1764	creation	creation
initiation	133	rite	creation
founding	396	creation	creation
foundation	1697	institution	school
origination	6	coorigination	creation
creation	1180	establishment	formation
colonization	228	colony	settlement
settlement	2649	institution	bank

Figure 5.10 Results of Harvard Experiment

specific instances of university names[9]. If it did, then we could reasonably assign *Harvard* to that same subtree. The difficulty lies in where to place *Harvard* in the absence of knowledge of the terms it is found closest to. If we ask SEXTANT to compare *Harvard* only to the words which were used to generate the corpus, we find that *Harvard* is closest to *academy*. *Academy* is in the correct subtree, but now we must ask, should *Harvard* be placed as a child of *academy*, on the same level as *academy* or somewhere else in the subtree? To complicate matters *academy* reappears as a child of the synset *school, educational institution*.

The results of this experiment are mixed. On one hand, the hypernyms of *institution* form a group despite 22,000 other candidates and the new hyponym *Harvard* falls into the correct subtree of *institution*. But, at the same time, *Harvard* is not closest to the word *university*, as we would have hoped. The reasons for this seem clear enough. First, as the corpus shows, *Harvard* appears not only as a *university*, but as a *law school*, a *college*, a *divinity school*, an *educational review*, and as a *business school*. Each one of these entities is an *institution*. Second, *Harvard* most often appears as a proper noun modifying one of these entities. It appears 23 times in the expression *Harvard University* and 19 times in *Harvard Law School*. In this case, it is closest to *Yale*, which appears 14 times as a *Yale University* as well as 10 times as *Yale Law School*. This suggests that proper names, which are treated exactly the same[10] as common nouns by SEXTANT, should actually be treated differently.

Suppose that there did exist a subtree containing *Yale, Cambridge, Columbia, ...*, we would most probably find *Harvard* already there. However, this experiment shows that SEXTANT places terms near similar terms, and this ability to place a new term near its closest lexical neighbors might be useful in non-static domains in which new terms are introduced over time (e.g., the medical domain). If a knowledge structure exists for the domain, new text will produce new terms that must be integrated into this structure. This experiment provides an indication that such integration may be automated using existing lexicons, grammars, and text corpora that cover the relevant phenomena well.

[9] The unabbreviated version of the network beneath *institution* does contain names of specific religious groups, although no specific names of universities appear.

[10] The preprocessor described in Appendix 1 allows for sequences of non-initial capitalized words to be joined into one unit. This preprocessor can be applied to normal mixed-case text. The resulting units are treated afterwards by SEXTANT exactly as common nouns. This preprocessor was not applied to the HARVARD corpus here so that the string *Harvard* appeared as a separate unit.

word	*freq*	*closest terms*
wheat	234	rice, corn
rice	217	wheat, corn
corn	182	rice, wheat
crop	179	wheat, cultivation, rice, grain, production
meal	129	corn grain food, export, product, rice, hay,
product	102	export, production, industry, cattle
grain	101	export, growing, plant, cereal, producer
production	93	cultivation, growing
area	90	region, farm, land
food	89	export, grain, crop
plant	84	variety grain, production, seed
center	79	distribution, production, product, export, farmer
cultivation	71	growing, production,
farming	65	field, farm, export
	...	
millet	59	francois, sorghum, barley
barley	46	potatoe, cotton, rye
oat	44	barley, vegetable, potato, cotton
cereal	41	producer, starch, flour,
buckwheat	10	bread, sugarcane, pineapple, sorgum
grit	10	wing, farina, estrilidae
loblolly	8	grade drought, pine
gruel	3	second, conglomerate
mush	3	cornmeal
oatmeal	3	quaker, gascony
porridge	3	symbol, ale, roll
cornmeal	1	counterpart, sports-writer, literature
farina	1	embayment, peel
pudding	1	NO RELATIONS

Figure 5.11 Results of Cereal Experiment

5.2.3 Other experiments

As another example, the hypernym acquisition algorithm discovered the lexical relation between *rice* and *cereal*. WordNet has two senses of *cereal*: one as a grain, and one as a breakfast product. We extracted 260,000 characters of text from *Grolier's* comprising sentences containing one of the following strings:

> frumenty, kasha, grits, hominy, grits, farina, millet, oatmeal, hasty pudding, mush, burgoo, flummery, gruel, loblolly, porridge, rice, oat, cornmeal, meal, corn, wheat, buckwheat, barley, cold cereal, puffed wheat, puffed rice, wheatflakes, cornflakes, granola.

In Figure 5.11 we see that *rice* is found to be closest to *wheat*, and vice versa. This table also shows the results involving the breakfast product terms (some did not occur in the corpus); note that the frequency of these terms is too low for valid assessments to be made. For example, the cold cereal terms that were unambiguous, such as *wheatflakes*, *cornflakes*, and *granola*, as well as all of the hot cereal items, are underrepresented in the corpus. This fact reduces the possibility that *rice* could be found similar to the breakfast product terms, although the fact that it is strongly related to the *wheat* sense does lend some validity to the result. This example highlights another difficulty in applying the statistically-based SEXTANT to this problem: underrepresentation of data in the corpus can make the choice of positioning less reliable.

A third relation that we considered is *hyponym(logarithm, calculation)*. WordNet records *logarithm* as a kind of *exponent* which in turn is a kind of *mathematical notation*. However, the sense of logarithm as a calculation should also be considered. The WordNet structure for *calculation* is given in Figure 5.12.

Although there is no common one-word term for computing a logarithm, there should be an entry for *logarithm* alongside *exponentiation* in the second subtree (or perhaps *taking a logarithm*). Our results found *logarithm* to be closest to *extrapolation* and *multiplication*, one term from each subtree, and thus eluded correct classification. (It is not clear, however, that the first subtree is entirely well-defined. Why is *integral* associated with figuring, as opposed to listing *integration* in the second subtree?)

This example shows that difficulties arise when the shades of differences within the human grouping of terms are subtle, fine-grained, or somewhat arbitrary. The original goal was to make coarser-level distinctions, but because

```
Calculation: (hyponyms)
calculation, computation, figuring, reckoning
        => extrapolation
        => interpolation
        => estimate, estimation
              =>  guess, guesswork, shot, dead reckoning
              =>  approximation
       => integral
              =>  indefinite integral
              =>  definite integral
calculation, computation
      => mathematical process, operation
             =>  differentiation
              => division
              => integration
              =>  multiplication
              =>  subtraction
              =>  summation, addition
              =>  exponentiation, involution
```

Figure 5.12 WordNet entry for *calculation*.

certain portions of WordNet are so well-developed, it turns out that many decisions require choosing between finely divided subtrees.

The previous three examples have shown various difficulties associated with this approach. Foremost among them is the fact that WordNet senses have been manually created and correspond to a human conception of what attributes of two words or concepts are saliently similar, whereas SEXTANT finds similarity based on frequency of usage in a specific corpus. It may well be that two concepts considered to be similar in the WordNet hierarchy do not display the kinds of regular contextual similarities that our methods can recognize in a particular corpus.

There are other difficulties as well. For example, in one instance we found the relation *hyponym(granite, rock)*. In WordNet there are very fine differences among the senses of *rock* and in fact *granite* appears in more than one of its subtrees. Other difficulties we found were: the hypernym senses have no child subtrees against which to compare, the hypernym is a very general term and thus has hundreds of children, and the hyponym does not appear frequently enough in the corpus to make statistical observation possible.

5.2.4 Summary

In this section we have described an attempt to combine three different text analysis tools: a human-built knowledge source, a statistical similarity measurement applied to syntactically derived contexts, and a pattern-based relation extractor. The lexical relations are used to augment the lexical hierarchy and the similarity measure is meant to determine the part of the hierarchy in which the relation belongs, while simultaneously the lexical hierarchy selects the appropriate part of the corpus to feed to the similarity measure. Our results point toward the possibility of placing words into an existing knowledge structure by using similarities to parts of that structure. The similarity calculations based upon shared lexical syntactic contexts often place words in the same classes in the man-made WordNet thesaurus. But the experiments described here point out that this effect is strongest for words possessing much context. The problem of placing rarely appearing words still remains unresolved, even given a lexical syntactic pattern that gives an explicit clue as to its relation with another word.

5.3 WORD MEANING CLUSTERING

The products of SEXTANT presented so far have been lists of words recognized as being similar to a particular word, given the syntactic contexts of the words being compared. But even within one coherent corpus, such as one treating medicine or one treating business, a word may have many nuances of meaning. Two words may be recognized as similar to a third word for different reasons, along different axes corresponding to nuances of meaning of that third word. For example, in a medical corpus, *administration* can relate to the organization of a hospital or to the injection of the drug. In this section we explore the possibility of using SEXTANT not only for determining which words are similar, but grouping these similar words together along semantic axes.

In order to define these axes, we modify a concept introduced in Hindle (1990) and define words as being ''reciprocally near neighbors'' if the words appear on each other's similarity lists within the closest N words (we use $N = 10$ throughout). These words can serve as seeds for axis definition in the following way. We consider the following case. Let us suppose that a word *A* was found close to *B*, *C*, *D*, *E*, and *F*, and suppose that *B* was reciprocally near to *A*; that is we suppose that *A* was also one of the closest words to *B*. We can be confident that *A*-*B* forms a semantic axis and try to attach the other words *C*, *D*, *E*, and *F* to this axis. One way to do this is to include any of these words which is also a near neighbor to *B*, independent of *A*. This defines a set of words which are (1) close to *A*, (2) near neighbors to *B*, and (3) close to this axis, supposing that *A*-*B* is a semantic axis.

When this grouping technique is applied to the most frequent words from the MERGERS corpus, we develop the clusters presented in Figures 5.13 and 5.14. Words are included in a cluster if they are as frequent as or more frequent than the second word defining the axis; taking into account frequency in this way is an attempt to generalize from more specific to more general words. For example, in the table below, we see that *agreement* is a reciprocal near neighbor to *acquisition* in the MERGERS corpus, so we take *acquisition-agreement* to be one semantic axis of the word *acquisition* for this corpus. Then, comparing the similarity lists of *agreement* to the other words closest to *acquisition*, we discover that *bid*, *offer*, and *plan* are more general words (appearing more often) than *agreement* and are reciprocally close to it. This group seems to define a sense of *acquisition* having to do with the negotiation process involved in acquiring some company.

Also in Figure 5.13, we can see similar sense differentiations in the *approval-action-decision* as opposed to the *approval-authority-review-rule* which distinguish the act of approving from the right of approving. On the other hand, we also see correspondences unlike ones that human would draw such as between *agency-United States-thrift*. Although these words are all connected through the *Resolution Trust Corporation*, it is difficult to see a clear semantic axis here. Let us look at the attributes that all three terms have in common, shown in Figure 5.15.

This observation does not clarify the situation further, since each attribute seems to be adding one little piece of meaning to the composite judgment made by SEXTANT that the words are similar. A further problem with the clustering method as it exists is that sometimes the distinctions seem to be too fine. For example, it might be perfectly satisfactory to group *acquisition-sale-purchase-transaction-merger* into one large group rather than its many small subsets as they appear in Figure 5.13. This level of distinction or grouping of course depends on the use to which these lists are to be put. If the use is for human consumption, such as an expansion proposing interface to a retrieval system (Nelson 1993), then larger groups would be all right, since the user could quickly pare down the list. If it is for an automatic system expansion system, then smaller lists might be preferable (Sparck Jones 1971).

When the same clustering technique is applied to the MED corpus, we get the clusters appearing in the Appendix (page 163) for words appearing more than 10 times in the corpus. Again, the technique of using reciprocal near neighbors creates axes which are able to group similar words, although the non-medical person must resort to an adequate medical dictionary such as *Taber's Cyclopedic Medical Dictionary* (Thomas 1985) to recognize the relations. For example, *a-crystallin* and *dna* are both examples of *proteins*. *Atresia* is a ''congenital absence or closing of a normal body opening,'' and should be close to the axis *abnormality-anomaly*. It is not clear what relation, if any, exists between *acid*, *fraction* and *protein*. *Acidosis*, though, is an *insufficiency* resulting from renal *hypertrophy* which is captured in the *acidosis-insufficiency-hypertrophy* axis in Figure 5.16.

Another interesting result, presented in Figure 5.16, is the way the word *tumor* is divided along malignant and non-malignant axes. One axis is *tumor-growth*, which attracts the words *tissue* and *effect*, while the axes *tumor-cancer*, *tumor-carcinoma*, and *tumor-lesion* bring in each other to their axes.

Semantic Axis	*words closest to axis*
acquisition *as an* agreement	bid offer plan
acquisition *as a* bid	offer sale
acquisition *as a* deal	transaction merger investment agreement
acquisition *as a* merger	transaction
acquisition *as a* plan	bid offer sale
acquisition *as a* purchase	bid transaction sale
acquisition *as a* sale	offer
acquisition *as a* transaction	bid offer sale plan
agency *as a* firm	bank concern
agency *as a* united-states	thrift
agreement *as an* acquisition	plan offer bid
agreement *as a* bid	offer
agreement *as a* deal	transaction acquisition investment
agreement *as a* plan	offer bid price
agreement *as a* proposal	plan offer bid transaction
airline *as a* carrier	government institution
airline *as an* institution	government
airline *as a* security	investor
analyst *as a* bank	group year
analyst *as a* executive	group company
analyst *as a* group	company year
analyst *as an* investor	bank concern firm
analyst *as an* official	executive
approval *as an* action	decision
approval *as an* authority	review rule
approval *as a* clearance	review authority
approval *as a* review	clearance authority
approval *as a* vote	step
asset *as a* control	part
asset *as an* investment	interest
asset *as an* operation	interest business
bank *as an* analyst	group
bank *as a* firm	group market
bank *as a* group	year
bank *as a* investor	firm analyst
bank *as a* market	year
bid *as an* acquisition	offer plan agreement sale
bid *as an* agreement	offer plan
bid *as a* plan	offer sale
bid *as a* proposal	offer plan transaction agreement

Figure 5.13 Semantic clusters from the MERGERS corpus.

Semantic Axis	*words closest to axis*
board *as a* chairman	executive spokesman
board *as a* director	shareholder chairman official
board *as a* management	shareholder executive
board *as an* offer	stock
board *as an* official	executive
board *as a* shareholder	stock
board *as a* spokesman	executive official
bond *as a* fund	debt security
bond *as a* loan	debt security fund
bond *as a* security	fund investor
bond *as a* thrift	loan
business *as an* asset	operation
business *as a* concern	market
business *as a* firm	concern market
business *as a* maker	concern
business *as an* operation	concern market
business *as a* sale	share
buy-out *as a* deal	transaction takeover merger acquisition investment
buy-out *as a* merger	transaction acquisition
buy-out *as a* purchase	transaction acquisition
buy-out *as a* takeover	merger investment
buy-out *as a* transaction	acquisition offer
. . . . .	. . .
transaction *as an* acquisition	offer plan sale bid
transaction *as a* bid	offer sale
transaction *as a* buy-out	purchase
transaction *as a* deal	buy-out merger acquisition
transaction *as a* merger	buy-out acquisition
transaction *as a* plan	offer sale bid
transaction *as a* proposal	offer plan bid
transaction *as a* purchase	acquisition sale bid
transaction *as a* sale	offer
value *as an* amount	cash number debt
value *as a* cash	profit debt
value *as an* earning	profit
value *as a* financing	cash
value *as an* interest	price
value *as a* price	year
value *as a* profit	price

Figure 5.14 Semantic clusters from the MERGERS corpus, cont.

acquire-DOBJ	agree-DOBJ	base-DOBJ	big	continue-SUBJ
create-DOBJ	expect-DOBJ	hold-SUBJ	lose-SUBJ	make-DOBJ
make-IOBJ	medium-size-DOBJ	new	own-SUBJ	plan-DOBJ
regional	report-SUBJ	say-DOBJ	say-SUBJ	second-large
sell-DOBJ	small	think-DOBJ	time	top year

Figure 5.15 Attributes shared by *agency*, *United States*, and *thrift* in the MERGERS corpus

Semantic Axis	*words closest to axis*
a-crystallin *as a* dna	protein
ability *as a* capacity	production function
ability *as a* inability	capacity
abnormality *as a* anomaly	atresia
abnormality *as a* impairment	disorder disturbance
abnormality *as a* nature	manifestation
absence *as a* sibling	family
absorption *as a* exchange	transport
absorption *as a* na	exchange
absorption *as a* po	tension
accumulation *as a* extent	jaundice
acid *as a* dna	protein
acid *as a* fraction	protein
acidosis *as a* insufficiency	hypertrophy
act *as a* prolongation	deficiency
activity *as a* amount	concentration level number
addition *as a* absence	presence
adenocarcinoma *as a* carcinoma	tumor
adenoma *as a* hyperplasia	hypertrophy
adjunct *as a* chemotherapy	therapy
administration *as a* dose	injection
administration *as a* infusion	dose
administration *as a* secretion	deficiency
administration *as a* therapy	treatment
administration *as a* treatment	response
.....	...
tumor *as a* cancer	lesion tissue
tumor *as a* carcinoma	cancer disease
tumor *as a* growth	tissue effect
tumor *as a* lesion	cancer disease
tumor *as a* tissue	disease
.....	...

Figure 5.16 Semantic clusters from the MED corpus.

cancer :: [255 contexts, frequency rank: 29] MED *Relat.* lesion, tumor; tissue, disease; carcinoma. *Vbs.* advance, disseminate. *Exp.* cancer patient (cf. survival time, joint deformity), cancer chemotherapy (cf. survival time, intra-arterial infusion), cancer cell (cf. human cell, year period). *Fam.* cancer-specific, cancerous.

Figure 5.17 Automatically extracted thesaurus entries for *cancer* from MED.

5.4 AUTOMATIC THESAURUS CONSTRUCTION

We have seen now that a number of different pieces of information can be extracted from a corpus, using the syntactic contexts of the words within it. We can extract:

- similar words (cf. Chapter 2)
- semantic axes around a given word (cf. Section 5.3)
- specific expressions containing the word (cf. Section 3.2.3)
- relations between these expressions (cf. page 146)
- families of words (cf. page 106)

We can use this information to create a first draft of a thesaurus entirely automatically from a given corpus. For example, reuniting all of this information from the 1MB corpus `MED` for the word *cancer* produces the thesaurus ''entry'' given in Figure 5.17.

The structure of this entry shows that, in this corpus, the word *cancer* possesses 255 contexts (attributes) by which to judge its similarity. This word *cancer* has the 29th greatest frequency for all the words compared, placing it near the top. The label *Relat.* shows that the word is found to be related to *lesion, tumor; tissue, disease; carcinoma* using the semantic axes techniques of Section 5.3. The semi-colons separate words appearing with about the same frequency from words appearing more frequently, and then from words appearing less frequently than *cancer* in this corpus. This comparative frequency has been used as a means of creating hierarchical relations between words in a corpus (Srinivasan 1992), though it is an extremely weak statistic and certainly better can be found. The *Vbs.* section lists the most frequent verbs with which this word associates as a subject, or as a direct or indirect object (cf. page 37). The section *Exp.* gives the most frequent noun phrases in which this word appears (cf. Section 3.2.3). This group is produced by extracting all noun phrases that were unambiguously structured in the corpus. Since we have not attacked the

cancer :: [905 contexts, frequency rank: 16] AIDS *Relat.* disease; failure, ascites, lesion, tumor, carcinoma. *Vbs.* advance, develop, treat, smoke, detect, use, review, randomize, induce, increase. *Exp.* cancer treatment (cf. median follow-up, hospital stay).

Figure 5.18 Automatically extracted thesaurus entries for *cancer* from AIDS.

problem of determining the internal structure of noun phrases in any depth (cf. Section 3.4.3), the unambiguous noun phrases recognized were limited to two-word noun phrases appearing above a certain threshold (here, three times). These phrases are followed by the phrases which are closest, using a technique described below on page 146. The family of words are words which appear frequently in the same document as the head word and which satisfy the morphological similarity criteria given on page 107. Unifying these categories of information together in one entry provides a credible first pass at a thesaurus entry for this word. At this point one should remember that neither human filtering nor semantic information entered into the production of this entry.

The word *cancer* appears frequently in one other of our corpora, in AIDS, a collection of recent abstracts on this illness. Producing the entry for *cancer* from this source gives the slightly different result in Figure 5.18.

It is interesting to compare the automatically extracted entry for the same word from two very different corpora. In Figure 5.19 we give the entries that our system produces for the word *growth*. The first entry is from `MED`, the corpus of medical abstracts, and the second is from `MERGERS`, a corpus of Wall Street Journal articles on mergers; one could guess the sources by the associations made by SEXTANT. The medical corpus generates relations between growths in the physical sense of a growth, while the mergers corpus associates growth with gains, losses and performances. It is also interesting to note that the expression *growth rate* is common in both corpora though associated with very different expressions. In the medical corpus, *growth rate* is associated with *growth retardation*, while in the financial corpus it is associated with *future performance* and *profit margins*.

This one example of the different corpus-dependent meanings of a relatively common English word demonstrates the potential of knowledge-poor methods, such as those described in this book, to alleviate the problems posed by *word variability*. Indeed, *word variability* is what has motivated this research.

growth :: [284 contexts, frequency rank: 25] MED *Relat.* tumor; effect, tissue; antigen, protein, development. *Vbs.* retard, stimulate, show, follow, enhance, accelerate. *Exp.* growth hormone (cf. bone marrow, parathyroid hormone), growth rate (cf. growth retardation, folic acid), *Exp.* tumor growth (cf. body growth, tenuazonic acid), growth retardation (cf. dna content, body weight), body growth (cf. tumor growth, body weight). *Fam.*

growth :: [320 contexts, frequency rank: 139] MERGERS *Relat.* level, increase, gain; loss; performance, return, rise, decline, flow, expansion. *Vbs.* say, expect, slow, accelerate, maintain, sustain, forecast, continue. *Exp.* rapid growth (cf. buy-out bid, raise capital), profit growth (cf. electronics group, total revenue), growth rate (cf. profit margin, future performance), growth potential (cf. company spokeswoman, board seat), future growth (cf. specialty chain, bottom line). *Fam.*

Figure 5.19 Automatically extracted thesaurus entries from two different corpora for the word *growth*.

Again, here is a word with many meanings in everyday English. Given a query using this term on a given collection of text, we could automatically suggest, using the results produced by SEXTANT, how this word is used in the collection, what it associates with, and give a more complete picture of the sense of the word than any stemming or key-word-in-context systems are capable of producing. If you consider that words and expressions can be indexed into documents using the traditional inverse indexing mechanism, it is not hard to see the utility of a suggestion-making interfaces, such as the one sketched here, as front-ends to any retrieval system.

A complete example of automatic thesaurus generation for the MED corpus can be found in the Appendix. Shorter extracts are found after each corpus description in the appendices.

5.5 DISCUSSION AND SUMMARY

In Chapter 3 we presented a selective natural processing technique which extracted lexical syntactical contexts of words in a corpus. A statistically-based similarity measurement scheme compared these contexts to create list of similar words. Chapter 4 provided a demonstration that these syntactically-derived lists capture semantic similarity. The current chapter provided some possible applications of this objectively specifiable similarity. The fundamental concern of this work is acquiring an understanding of corpus-dependent word

usage. Knowing how the words in a corpus (in a subdomain specified by a collection of natural language text) are used is necessary for understanding that domain. We present here four different applications of the information gleaned from compared word usage.

For information retrieval, these lists provide an idea of what subjects are discussed in the document base. If the most common words in the document base are considered as the characteristic vocabulary of that corpus, SEXTANT provides a clear image of how this characteristic vocabulary is used. Our practical experiments show that automatically augmenting user queries with the most similar words can sometime improve the precision of the documents returned by a traditional retrieval system. The results do not conclusively prove the usefulness of this automatic expansion, we argue, since often what users want to find are distinctions among similar objects. There are two ways of using these similarity lists.

1. As propositions to the user in a classical information retrieval system, as we discussed in Section 5.1. The user can then know that the words in his or her query have certain associations within the corpus that he or she is interrogating. If the user accepts certain of these suggestions for expansion, then his or her results can become more precise, as shown in Figure 5.6.
2. In the 'evoking' stage of more modern retrieval systems (Evans *et al.* 1993; Cutting *et al.* 1992). In large textual bases, these systems perform a quick pass to gather candidate documents which are then subjected to more refined and computationally expensive discriminating phase before being presented to the end user. Incorporating words similar to the query terms, such as SEXTANT can provide, would allow a certain quantity of silence, i.e., documents missed because the words used in the query are not those in the document, to be eliminated.

A second application is placement of a new word in an existing knowledge structure. Experiments with WordNet pointed out possibilities and weakness of both this resource and our proposed application. The weakness of our method for this application is the general weakness that has shown up throughout this book, viz., the syntactic-based comparison technique requires an adequate quantity of context in order to make accurate judgments. Using another knowledge-poor technique which proposed hypernym-hyponym pairs, we tried to apply our method to decide in which subtree under the hypernym to place its hyponym. Although SEXTANT often found the word as being most similar to other words in the proper subtree, there was rarely enough context

to decide against other possible placements, the other subtrees corresponding to concepts poorly represented in the corpora that we tested. The weakness of WordNet which came to fore was a certain confusion between senses; some synonym sets were very similar to others, their separation based on unclear distinctions. This is a problem that will be apparent in any man-made information source made for human consumption that can rely on presupposed semantic knowledge in the human reader to make clear distinctions.

The third use of the information generated by SEXTANT finds its application in semantics as the study of meaning. If two similar words are allowed to define an axis of meaning, other words can be placed along that axis. We have shown that the information that SEXTANT extracts allows the creation of these axes. In other words, general senses of meanings of words can be derived from their lexical syntactic uses in a corpus, and we have a means of objectively recognizing and explaining these senses. Although the only application we provide here is the production of theses axes, we feel that this ability to produce these axes is the first step toward the automatic creation of corpus-induced semantic markers. For thirty years, since Katz's and Fodor's first description of semantic markers and their necessity for text understanding, it has always been presupposed that semantic markers are given, somehow encoded in the lexicon. The only large scale attempt to encode semantic markers on a general basis (Lenat *et al.* 1986) has demonstrated that the task is difficult to realize by hand. Automatic creation of domain-directed semantic markers would have a great number of applications in artificial intelligence, from text understanding systems to machine translation.

The fourth application is the most direct application of SEXTANT's methods. The creation of a first pass of a thesaurus in a completely automated fashion is presented in Section 5.4. The thesaurus presented is a second order thesaurus providing relations between semantically related but orthographically different words and terms. As in all of the preceding applications, the results produced are most informative for the characteristic words of the corpus. Results for rare words are sporadic and noise-filled. Taking this caveat into account, the thesaurus produced must be considered only as a rough sketch. Nonetheless, it is a base to work on and improve. It is the most exciting result produced by SEXTANT since it promises that, what before was a completely manual and fastidious task, can be partially automated.

6

CONCLUSION

6.1 SUMMARY

In the preceding chapters we have examined a selective natural language processing approach to extracting corpus-specific semantics. We described the motivation of this approach: the *language variability* problem that affects any computer-based manipulation of text, e.g., information retrieval, filtering, language understanding, human-computer interfaces, machine translation. This problem generated much research in computer-based semantics, a portion of which we reviewed before we presented our system SEXTANT.

SEXTANT processes raw text in the following manner. Given a corpus of text divided into documents, our system tokenizes the text into word units, performs morphological analysis, looks up each word in a lexicon, disambiguates the grammatical category of each word[1], parses the disambiguated text into noun phrases and verb phrases, and creates the dependencies between the words within these phrases and between the head words of phrases in the same sentence. These dependency relations are then considered as attributes of the words involved. These attributes form the recognized context of a word in SEXTANT. The attributes are compared by a module implementing a weighted Jaccard similarity measure in order to discover words used in a similar manner throughout the corpus. One preliminary result produced by SEXTANT is a list of similar words for each word in the corpus.

[1] These last three steps (morphological analysis, lexicon look-up, and grammatical disambiguation) are performed by programs that were written by others at the Laboratory of Computational Linguistics, Carnegie Mellon University. We have used them in SEXTANT with permission of the director of this laboratory. All other steps of SEXTANT's processing were created and developed by the author and are independent of these modules.

The details of this processing and sample results were reported and discussed. We also discussed the complexity of SEXTANT processing steps in function of the vocabulary size of the corpus treated. We showed that the results produced by SEXTANT became more stable as more context was added for a word.

In order to evaluate the plausibility of our claim that we were extracting semantically similar words, we performed three distinct evaluations of SEXTANT. We showed that the results of its association techniques correspond to psychological data from word association tests. We created artificial synonyms and then demonstrated SEXTANT abilities to recognize them. Finally, we measured the results produced by SEXTANT against a series of gold standards, calculating their overlap. In this last evaluation stage, we also showed that our syntactically-based technique produces a greater overlap with existing thesauri than classic windowing techniques for the characteristic vocabulary of the corpora tested.

We presented some potential applications of the similarity discovery techniques of SEXTANT. We implemented a query expansion scheme that added in words found to be similar to query terms, and tested it on classic information retrieval testbeds. The results were mixed but showed that in an interactive setting, the automatically generated expansion propositions can improve precision of the documents recalled. We then showed how SEXTANT can be combined with another knowledge-poor technique in order to enrich an existing thesaurus. Then more general linguistic information was educed from the preliminary processing of SEXTANT. We showed that closer examination of the causes for similarity decisions allowed us to create clusters of words along semantic axes, a first step in the automatic creation of corpus-derived semantic markers. Finally, we showed how the natural language processing techniques developed in this book can be used to produce a credible first pass at a corpus-based thesaurus directly from its raw text.

We conclude from our work that syntactically derived information can form a basis for discovering semantics between words. As an added interest, when the corpus from which the information is derived defines a domain, the relations discovered are specific to that domain. The results presented here, and in the appendices, demonstrate the real possibility of automating the semantic processing needed for artificial intelligence approaches to natural language. As a subsidiary result, we also have found that classic techniques such as using textual windows as contexts of words, although they provide less accurate results that the syntactic contexts used by SEXTANT for the more common

words in the corpus, can be used to advantage for rare words for which little syntactic context exists.

6.2 CRITICISMS

6.2.1 Parser Problems

Since SEXTANT uses syntactically derived information to determine similarity between words, one would hope that this information be correct. If it is not correct, then SEXTANT defaults to a form of classic textual windowing in that words are joined together simply because they are in the same general location. As an example, this criticism can be leveled against our treatment of progressive verbs. As mentioned in Section 3.7, the recognition of progressives as gerunds or as adjectives is a difficult, and probably domain-dependent problem. In this case, as with other problematic parsing problems such as prepositional phrase attachment, we have chosen a quick an easy solution based on our intuition of the most common cases. We defend ourselves weakly by claiming that such heuristic choice-induced errors are at least regular, so that if a certain textual pattern repeats, the derived context by which similarity calculations are made will also be regular in its 'error,' and the matches between words in similar contexts will still be made.

A more serious parser problem that we have discovered throughout our experiments is the unmet need for distinguishing between certain types of nouns. Although we wish to eschew a preliminary semantic typing of nouns, there are a limited number of word types that can be recognized and should be distinguished from common nouns. Examples of these types are proper names, place names, expressions of time, dates and measures. As we have presented a simple name recognizer (see Appendix 1) so we need a means of recognizing dates, times and measures. A gazetteer should exist alongside the general English lexicon for place names. And once these types have been recognized they should be extracted from similarity comparison with other types of nouns, something which has not been done here. A workshop (Boguraev 1993) has been held on the recognition and use of such entities, which have been lumped together with common nouns and parsed and compared as such.

6.2.2 Unobserved Clues

We have been using primarily one source of information about the possible meaning of a word, its lexical-syntactic attributes throughout a corpus. But we have ignored other explicit clues about a words meaning in a corpus. If we present an automatic thesaurus generation technique these clues should be taken into account. For example, we have mentioned the technique developed by Hearst (1992) for recognizing hypernymic relations from explicit lexical patterns obeying certain restrictions. She was able to use this technique to extract 330 such relations from 50 Mbytes of encyclopedia text. If such proportions were to hold for our tested corpora, we should expect to find from 7 to 30 such hyponymic relations in each corpus.

Other explicit clues that we might have used to enrich our thesauri are morphological variations. Although we present a technique in Section 5.1.4 for recognizing a portion of these morphological variations, our approach is rudimentary. A better approach would have been to apply a suffix and prefix analyzer and then compare document contexts, thus providing a larger class of corpus-derived family words than our simple stem matcher provides.

There are other clues that we ignored and that are less explicit but which may yield more, useful contextual information about word meaning. For example, we ignored all pronouns and other anaphoric references. In the `AI` corpus, 2.8 Mbytes of artificial intelligence abstracts, there are 3501 words[2] tagged as a pronoun among the 387,000 words comprising the corpus. There are 25,000 unique words in this corpus; if the pronouns were resolved equally over all these words, then only one word in five would possess one additional attribute. More probable though is that anaphoric reference is distributed according to Zipf's law (see Figure 4.4, page 4.4) so that resolution of anaphora would apply to words according to their frequency. If this the case, then since pronouns account for 1% of the total words, we can suppose that resolving anaphora would augment each word's contexts by about 1% also.

Another potential resource for judging word meaning that we have ignored is the part of the document in which the word is found. Finding a word in a list or in a table would provide a key as to its meaning. In addition, knowing whether the word appears in the introduction, body, or conclusion of a document would certainly give an idea of its importance, more than the simple weighting scheme that we employed in Section 3.2.6. Our techniques are applied in a

[2] This includes all occurrences of the word 'it,' even non-anaphoric uses such as in ''It is clear . . . ''

blindly equal way to the whole document without regard to structure. This approach was chosen for facility but could certainly be improved.

6.3 FUTURE DIRECTIONS

6.3.1 Infrequently Appearing Words

When a word or term appears only a few times in a corpus, the fine-grained syntactic analysis of SEXTANT does not provide enough context to judge the meaning. In some cases, for example when the word appears in a recognizable list (Hearst 1992), clues are explicitly given. In the absence of such explicit clues, wider contexts around the unknown word must be used, such as words appearing in the same sentence, paragraph, or document. Crouch (1990) reported some success in improving information retrieval by associating rarely occurring words from similar documents[3]. We believe that some of the unobserved clues mentioned or more extensive use of recognizable semantic contexts (Robison 1970) may yield reliable associations for these words.

6.3.2 Semantic Axes

Though we have suggested in Section 5.3 that semantic axes might be extracted from the syntactic contexts we use, it is not always clear how to name the relations in these axes. It would be interesting to be able to specify which kind of context produces which kind of relation. Sometimes relations that SEXTANT extracts seem to be somewhat synonymous in the textual domain, as in the examples of extracted relations from MED in Figure 6.1. Sometimes the relations seem to be part-whole, as with the examples in Figure 6.2. Figure 6.3 shows examples of words with the same level of genericity. Sometimes the relationship is difficult to express, see the examples in Figure 6.4.

We suspect that the data that we use for recognizing similarity is not sufficiently powerful for characterizing the relations between the similar words, and that some other information present in a large corpus might be used. For example, it is quite possible that true hypernyms, e.g., *dog* and *collie*, would share the same syntactic relations and be considered similar by SEXTANT. In order to

[3] Although when we asked her what these associated clusters looked like, if the pairings made sense, she replied that she had never looked at them, that her system was used as a black box.

case patient	response effect	treatment therapy
level concentration	method technique	increase rise
form type	cancer carcinoma	

Figure 6.1 Examples of extraction of near-synonyms.

cell tissue	change increase	child patient
development change	disease lesion	tumor carcinoma

Figure 6.2 Examples of extraction of part-whole relations.

day hour	rat mouse	dog rat
woman male	month yr	host parasite
artery vein	female male	salt ion

Figure 6.3 Examples of extraction of genericity relations.

factor role	acid concentration	process mechanism
area structure	diagnosis etiology	analysis investigation

Figure 6.4 Examples of extraction of relations which are difficult to classify.

conversive	$Conv$(to buy) = to sell
noun-verb	S_0(to move) = movement
adjective-noun	A_0(sun) = solar
verb-noun	V_0(death) = to die
noun-mode	S_{mod}(to write) = handwriting
noun-instrument	S_{instr}(to think) = brain
single instance	$Sing$(news) = item
noun-result	S_{res}(to hunt) = bag
...	...
verb-permit	$Perm$(to fall) = to drop
noun-incipient	$Incip$(war) = to break out

Figure 6.5 Samples of specific relation types defined by Apresyan and Mel'cuk.

determine that one word is more general than another, other clues must be exploited, such as presence in the same list of items as used by Hearst (1992). However, this difficulty of clarifying the relations between words might be hard not just for an automatic system but for human beings also. Chaffin & Herrmann (1988) explain that although some relations such as synonymy and genericity-specificity are generally well understood, and even actively taught in grammar school, recognition of other types of relations is a creative process. They write that

> Relations vary in the ease with which they can be expressed. Some relations require only a short phrase, e.g., *part of*, others require more elaborate expression, e.g., *a component that produces sound by striking* ... individual instances of relations may be seen as examples of more than one relational concept, e.g., *exhibit-display* ... a new relation may be recognized as an elaboration or concatenation of other, familiar relations. [p. 331]

In lexicographical work (Apresyan *et al.* 1970), a number of types of semantic relations have been proposed. For example, in addition to the classic relations of synonym and antonym, generic and specific, Apresyan and Mel'cuk propose a long list of possible types of relations between words, such as those given in Figure 6.5.

Fox (1980) used a reduced version of this typology in order to manually derive a semantic classification of words in an information retrieval setting. It would be interesting to see how many of these relations might automatically be extracted and by what automatically recognizable contexts. Once systems

SAMPLE TEXT	HYPOTHETICAL MULTI-WORD CONTEXTS	ONE-WORD CONTEXTS
Prosecutors dropped an illegal arms case against a retired Israeli army general and 10 other people after the defendants argued that the United States wouldn't have frowned on $2 billion in U.S. weapons sales to Iran.	prosecutor drop-SUBJ arm illegal case illegal case arm case drop-DOBJ israeli retired general retired army israeli general israeli general army case general defendant argue-SUBJ united-states frown-SUBJ weapon united-states sale united-states sale weapon sale iran	case illegal-arms illegal-arms illegal arms-case drop-DOBJ case army-general army-general army case israeli-general israeli-general israeli army-general israeli army-general retired general israeli-army israeli-army israeli weapon-sales weapon weapon-sales iran weapon-sales united-states

Figure 6.6 Given the above sample text, what our word-oriented extracts, and what is missed

such as SEXTANT have been applied to gigabytes of text, there might exist sufficient context to discover not only the similarity relations that we have seen in the previous chapters but also for typing the relations more accurately.

6.3.3 Multi-Word Phrases

During the course of this work, we have become convinced that restricting our work to individual words, while useful since many concepts are expressed as individual words, neglects a large portion of domain-dependent concepts that are expressed as multi-word terms. Were I to begin the work today, I would have used a larger representation of the textual units that would allow comparison between these phrases as well as between individual words.

For example, in Figure 6.6 we see the contexts that our word-based approach extracted for the given phrase, followed by the contexts that might be retained for a phrase-based approach. All of the relations of the word-based approach would appear in the phrase-based approach, plus these phrasal contexts. We

believe that this would supply richer results from a corpus, but, in order for this extraction to occur, the resolution of the internal structure of noun phrases must be solved. This internal structure is corpus dependent and probably decipherable from information present in the corpus, but the proper resolution of the problem would require further extensive research and testing.

We have already begun to explore what may be gleaned from considering two-word units. As we have stated, the proper way to proceed in this research is to determine what the basic units in a given corpus are, be they individual words or multi-word terms. This might be done by finding a means of recognizing unambiguous units which appear frequently without any intervening words, for example the unit *hot dog* will never appear with intervening words as in *hot brown dog*. Of course, since *hot brown dog* might appear in a corpus, the appearance of one intervening word is too strict a condition for rejecting a unit. Not willing to attack this task of discovering the basic vocabulary of a corpus, which would be a work meriting independent research in itself, we approximated such a discovery by extracting from each corpus all two-word noun phrases which appeared alone above a certain frequency. Examples of such frequent noun phrases are given in Figure 6.7.

As a preliminary experiment using more-than-one-word terms, we took all of these unambiguous two-word phrases, discovered by the method described in Section 3.2.3, and appearing more than five times in each corpus, and compared their contexts. Since the frequency of such phrases is lower than those of individual words, we extracted a wider context than the syntactic contexts that were used by SEXTANT in the first few chapters. This approach is suggested by the results we obtained in Section 4.5 in which a larger context, there the entire document which was used to calculate co-occurrence, produced better results for rare words than the finer-grained syntactic contexts. The contexts that we extracted in this experiment with two-word noun phrases were all of the other verbs, nouns and adjectives appearing within ten words before or ten words after the noun phrase, within the same sentence. Preliminary results, given below in Figure 6.8 and in the Appendix after each corpus description, are encouraging since they indicate that significant relations between longer terms can be discovered and that a combination of contexts - syntactic, windows and others - may be usefully overlapped and exploited.

MED	*AIDS*	*NEJM*
growth hormone	breast cancer	hiv infection
bone marrow	multiple myeloma	p24 antigen
breast cancer	blood pressure	hiv-1 infection
parathyroid hormone	heart rate	placebo group
blood pressure	side effect	cd-4 count
remain kidney	mg kg	base line
oxygen tension	mycosis fungoides	immunodeficiency syndrome
electron microscopy	blood flow	cell count
lymphoid cell	valve prolapse	t cell
dna synthesis	barrett esophagus	treatment group

JFK	*ANIMALS*	*MERGERS*
zapruder film	pit viper	executive officer
oswald rifle	body temperature	new-york-stock-exchange composite
jfk assassination	km h	tender offer
floor window	breed season	vice president
oswald guilt	hind leg	net income
news media	gestation period	hold company
president head	body cavity	joint venture
kennedy assassination	order passeriformes	takeover bid
jfk case	plant material	junk bond
cia agent	marine water	investment banker

Figure 6.7 Most Common Unmodified Two-Word Terms from Some Corpora

2-Word term [Contexts]		Groups of closest terms
blood-pressure [1745]	AIDS	heart-rate | blood-flow risk-factor heart-disease
risk-factor [995]	AIDS	heart-disease | blood-pressure blood-transfusion
valve-prolapse [906]	AIDS	panic-attack | chest-pain | heart-disease
control-group [899]	AIDS	treatment-group | placebo-group
median-survival [435]	AIDS	response-rate | survival-time
maintenance-therapy [147]	AIDS	remission-duration | median-duration
strike-zone [318]	BALL	power-zone | strike-call home-run | home-plate
regular-season [121]	BALL	post-season pennant-race
barry-bond [104]	BALL	bobby-bonilla post-season | strike-zone van-slyke
flag-incident [75]	BALL	color-guard | flag-upside
growth-hormone [1517]	MED	bone-marrow parathyroid-hormone growth-retardation
blood-flow [233]	MED	carbon-dioxide fluid-po2 | stroke-volume
radiation-therapy [138]	MED	survival-rate cancer-chemotherapy
pressure-curve [86]	MED	right-ventricle | left-ventricle stroke-volume
hormone-therapy [58]	MED	intra-arterial-infusion | steroid-therapy
hiv-disease [380]	NEJM	aids-related-complex immunodeficiency-syndrome
maintenance-therapy [263]	NEJM	initial-treatment amphotericin-b side-effect
heterosexual-contact [90]	NEJM	hiv-infected-person study-entry drug-abuser
study-design [89]	NEJM	study-medication | review-board study-subject
head-circumference [87]	NEJM	birth-weight | drug-addicted-mother

Figure 6.8 Samples of related two-word terms using window contexts

6.4 VISION

What we envision as a result of work like this research is the possibility of creating information systems that can use the topical clues of word presence in a document collection in a much more mobile, user-directed fashion than current text-based retrieval systems. The information seeker (Schamber *et al.* 1990) usually has only a vague notion of what he or she wants. If the system establishes one fixed way of representing its manipulable knowledge and this does not correspond to the user's need, what can he or she do but struggle to discover the system's structure?

What we would like to see is systems that adapt to the user's viewpoint via user-supplied clues as to his or her cognitive orientation.

Recently much more interesting and creative approaches to integrating the user into the retrieval process have been taken. A good example is the visualization interface proposed by Korfhage (1991). This system, in addition to offering classical information retrieval functionalities, provides a visual interface in which the user can position the words of his or her query on a two-dimensional plane; the documents corresponding to the words appear inside the polygon formed by the words, visually illustrating the pull of each word on each document. The user may reposition the words interactively and the documents readjust themselves to this new configuration. Such a system gives the user control. It allows the user to map his or her own vision of the spatial relations between information objects into the image space produced by the computer system (Treu 1990).

We feel that corpus-derived thesauri that result from our research will allow the same user oriented flexibility. We foresee that the user will be able to control the way that the system relates words by changing, dynamically creating, and overlaying different thesauri corresponding to different language uses.

For example, if a person were interested in John McCarthy's approach to AI, he or she could start with a selection of McCarthy articles, which would generate the most specific thesaurus. Word relations not found there could be sought from a larger (e.g., AI) thesaurus. The AI thesaurus, in turn, could be supplemented by a Computer Science thesaurus. All of these would be completed be a general English Language thesaurus.

Such a hierarchy could be used as an interface to a standard information retrieval system, as suggesting possible expansions for query terms. Were the

user more interested in, say, the economics of some computer system, he or she could use a hierarchy of thesauri derived from Economics, Computer Science, and General English. If the visual interface were something like a bookshelf metaphor as described in (Grefenstette 1991), the hierarchy of thesauri could be represented as a stack reference books, that would appear on an upper shelf, as over a work desk, and these thesauri would supply the lexical relations with which the user's queries would be interpreted.

In a multi-paradigm visualization system (Chang *et al.* 1991), the existence of various viewpoints on the knowledge present in the data, corresponding to different thesauri, would allow the research of information to proceed in many different ways, all presented concurrently on the same screen.

This vision presupposes the existence of both a pool of existing thesauri that the user could draw from, as well as a pool of online documents that the user could manipulate to create a new thesaurus tailored to his or her current research angle. The research that we have begun here will allow the creation of such thesauri.

1

PREPROCESSORS

LEX program for dividing text into words

```
Spaces        [ \n\t]*
Separator     (\'\'|\'\')
SentSeparator (\!|\?|\.)
Acronym       ([A-Za-z]\.([A-Za-z]\.)+|[A-Z]\.|[A-Z][bcdfghj-np-
tvxz]+\.)
Contractions  ('S|'D|'M|'LL|'RE|'VE|'s|'d|'m|'ll|'re|'ve)
Abbr1         (Co|Corp|vs)
Abbr2         (Jan|Feb|Mar|Apr|Jun|Jul|Aug|Sep|Sept|Oct|Nov|Dec)
Abbr3         (ed|eds|repr|trans|vol|vols|rev|est|b|m|bur|d|r)
%%
[ \t]*\n([ \t]*\n)+     { printf("\n\\$\n");
               /* Double new-line  read as a sentence break */ }

{Abbr1}\.               { ECHO ; printf("\n"); }
{Abbr2}\.               { ECHO ; printf("\n"); }
{Abbr3}\.               { ECHO ; printf("\n"); }

[0-9]+(\/[0-9]+)+       { ECHO ; printf("\n"); /*date*/ }
([+\-])?[0-9]+(\.[0-9]+|[0-9]+)?\%  {
                          ECHO ; printf("\n"); /*percent*/ }
(\$)?([0-9](\,[0-9]+)?)+(\.[0-9]+|[0-9]+)?    {
                          ECHO ; printf("\n"); }
[A-Za-z0-9][A-Za-z0-9]*((-|&)[A-Za-z0-9]+)*(s')? {
                          ECHO ; printf("\n"); }

{Acronym}              { ECHO ; printf("\n"); }
{Spaces}               ;
{Separator}            { ECHO ; printf("\n"); }
{Contractions}         { ECHO ; printf("\n"); }

{SentSeparator}({Spaces}|{SentSeparator})*  {
                     /* unambiguous SentSeparators */
```

```
                        printf("%c\n\\$\n", yytext[0]); }

"--"("-")*              { ECHO ; printf("\n"); }
"="+                    { ECHO ; printf("\n"); }
.                       { ECHO ; printf("\n"); }
```

AWK program for joining names

```
BEGIN           { InName = 0 ; Period = 0; }
/^ *$/          { next ; }
$0 ~ /^\\/      { if (InName==1) printf("\n");
                  Period = 1;
                  InName = 0; print ; next ; }
$0 == "."       { if (InName==1) printf("\n");
                  Period = 1;
                  InName = 0; print ; next ; }
$0 != "&" && ($0 ~ /^[^A-Z]/||$0 ~ /'s$/ || $0 ~ /s'$/) {
                  if (InName==1) printf("\n");
                  Period = 0;
                  InName = 0; print ; next ; }
                { if ((Period==1) && ($0 !~ /\./))
                        {print;}
                  else
                        {if (InName==0)
                           { InName = 1 ;
                             printf("%s",$0); }
                        else
                             printf("&%s",$0); }
                   Period = 0; next;
                 }
```

2

WEBSTER STOPWORD LIST

abaft
across
again
alone
although
anent
anywhere
as-for
astraddle
away
become
being
better
body
certain
could
downwards
eighteenth
eleventh
every
excepting
fifteen
five
forth
fourth
genus
have
herein
howbeit
inclusive-of
into
kept
least
make
million
mostly
near
next
nobody
novel
only
otherwise
over
percent
previous-to
rather
resembling
second
self
seventieth
shall
sixth
some
sometimes
syne
them
thereby
third
thirty-three
though
throughout
together
twelve
twenty-two
underneath
upto
wanting
whatever
whereas
which
whose
word
yourselves

aboard
action
against
along
always
another
apart
as-of
astride
back
becomes
below
between
both
chiefly
despite
due-to
eighth
else
everybody
exclusive-of
fifteenth
for-example
forty
frae
gets
having
hereupon
however
indeed
inward
kind
less
many
millionth
much
neath
nine
non-obstante
nowhere
onto
ought
owing-to
perhaps
prob
really
respecting
secondly
selves
seventy
should
sixtieth
somebody
somewhat
tenth
themselves
therefore
thirteen
thirty-two
thousand
thru
touching
twentieth
twentyseven
unless
used
water
when
whereby
while
will
would
zero

about
afore
agin
alongside
amid
anybody
apropos
as-to
astride-of
barring
becoming
beneath
betwixt
brit
circa
does
during
eightieth
elsewhere
everyone
failing
fifth
forby
forty-five
from
gone
hence
hers
implies
inner
irrespective-of
large
lest
marked
minus
must
neither
nineteen
none
off-of
opposite
ours
part
person
probably
regarding
respectively
seem
seven
seventy-eight
since
sixty
somehow
somewhere
than
then
therein
thirteenth
this
thousandth
thus
toward
twenty
twice
unlike
various
well
whence
wherein
whither
with
years

above
afoul
akin
already
amidst
anyhow
apropos-of
aside
athwart
became
been
beside
beyond
cannot
concerning
doing
each
eighty
enough
everything
family
fiftieth
form
four
further
group
here
herself
inasmuch
insofar
itself
last
light
meanwhile
more
myself
ness
nineteenth
noone
often
origin
ourselves
particular
place
process
regardless-of
said
seemed
seventeen
seventy-five
sith
sixty-nine
someone
specif
that
thence
thereupon
thirtieth
thorough
three
time
towards
twenty-four
twixt
until
versus
went
whenever
whereupon
whoever
withal
your

according-to
after
almost
also
among
anyone
archaic
aside-from
atop
because
before
besides
billion
cant
consequently
done
eight
either
even
everywhere
fewer
fifty
former
fourteen
furthermore
half
hereafter
himself
include
instead
just
latter
like
ment
moreover
namely
never
ninetieth
nothing
once
other
outside
particularly
please
quality
relating
same
seeming
seventeenth
several
sixteen
slang
something
state
their
there
these
thirty
thoroughly
thro'
times
tween
twenty-one
two-thirds
unto
very
were
where
wherever
whole
within
yours

accordingly
afterwards
aloft
alter
amongst
anything
around
aslant
attrib
because-of
beforehand
best
billionth
cause
considering
down
eighteen
eleven
ever
except
fewest
first
formerly
fourteenth
gainst
hardly
hereby
hither
including
instead-of
keep
latterly
made
might
most
naught
nevertheless
ninety
notwithstanding
ones
others
outside-of
pending
plus
quite
relatively
sans
seems
seventh
sextillion
sixteenth
small
sometime
such
theirs
thereafter
they
thirty-eight
those
through
tion
twelfth
twenty-twenty
under
upon
vis-a-vis
what
whereafter
whether
whom
without
yourself

3

SIMILARITY LIST

Here in order of decreasing frequency are the words extracted as similar by our initial investigation over the MED testbed. 1097 words were extracted and compared in all. For reasons of space, we present here all words appearing more than 100 times, plus samples of words appearing less frequently.

The first column is the word being considered, followed by its the number of contexts found for it in the collection. This is followed by the words calculated as closest to it, from left to right. Words having about the same similarity are grouped together. For example, the words *culture*, *tumor*, and *change* were about the same distance from *cell*.

word [*Contexts*]	*Groups of closest words*
cell [1156]	tissue | group effect patient study change level case activity
patient [883]	case | child group | treatment result study day | effect disease
effect [650]	change response | level action activity result increase study
study [626]	change | observation case effect patient result response rate
case [572]	patient | study lesion type child disease treatment result
change [549]	increase study effect | response difference | decrease pattern
level [548]	concentration | value rate excretion effect content | increase
acid [486]	protein | activity fraction dna increase glucose ratio value
result [446]	effect response observation patient study finding group data
child [412]	patient infant | group case subject | form woman year
activity [410]	effect concentration | increase level number response content
disease [401]	lesion | case change carcinoma | patient result type response
group [397]	patient | child result difference case | subject level day
response [389]	increase effect change | result reaction rate study treatment
rate [387]	increase concentration level | response value time result
increase [385]	decrease rise change | response reduction rate | difference
hormone [365]	serum | protein antigen dna | thyroid extract effect
tissue [350]	cell | growth cancer liver tumor | resistance disease lens serum
treatment [341]	therapy | patient administration case response | result effect
concentration [339]	level content | excretion value | rate ratio metabolism synthesis
defect [338]	disturbance case malformation | regurgitation type response
rat [331]	animal mouse | dog mice level | infant kidney day rabbit group
method [298]	technique | procedure test | mean result study | group treatment
pressure [286]	flow | volume | artery obstruction rate tension | serum sinus level
growth [284]	tumor | tissue increase effect development protein response
test [284]	technique method reaction | response study therapy observation
tumor [260]	carcinoma growth | cancer lesion | sarcoma tissue effect
blood [258]	level | tension concentration oxygen serum plasma liver | value
lesion [258]	disease | case cancer tumor symptom change | manifestation respose
therapy [256]	treatment | administration | drug response chemotherapy operation
cancer [255]	carcinoma | tumor lesion tissue | disease extract amyloidosis
type [249]	form case | change line feature pattern group defect disease syndrome
development [248]	growth change increase incidence production | response case pattern
reaction [245]	response test | effect increase relationship growth difference
factor [236]	role | mechanism difference change defect aspect treatment component
period [227]	time stage | group level result course duration change rate degree
difference [216]	change characteristic increase | rise correlation pattern decrease
content [212]	concentration | metabolism composition | fraction ratio synthesis
protein [212]	antigen | dna hormone growth acid analysis | concentration activity
culture [208]	marrow | suspension extract lung serum antigen kidney | specimen
syndrome [206]	type psychosis case lesion symptom disease result group treatment
injection [205]	administration dose | concentration time number response level
time [204]	day rate period | age serum incidence injection group month level
day [203]	hr hour | month week | year time patient group rat yr
value [202]	concentration level | increase rate decrease rise content | response
form [198]	type | case child sign | change problem result patient finding disease
fraction [196]	content | lens concentration antigen | serum preparation acid
dna [193]	protein | antigen hormone | mixture fraction a-crystallin polymerase
marrow [189]	liver | spleen serum suspension age kidney culture | experiment mice
technique [188]	method | test procedure | analysis change data dog therapy study

MED

word [*Contexts*]	*Groups of closest words*
antigen [184]	antibody protein | dna hormone component fraction growth
excretion [184]	concentration amount | clearance level | reabsorption retention
number [183]	concentration increase activity amount | response growth group
observation [181]	study result | finding analysis evaluation difference pattern
volume [180]	output flow pressure | concentration incidence detection ca
strain [178]	culture virus | type phage growth line species bacteriophage
evidence [177]	feature finding | study case information data severity cause
flow [177]	pressure | volume ph | circulation artery intake concentration
function [176]	capacity response disorder alteration relationship growth | stem
weight [172]	incidence content | concentration size hypertrophy ratio rate
pattern [171]	difference change | characteristic response feature observation
action [166]	effect | response growth | ability activity type result
synthesis [162]	concentration content metabolism | molecule production transport
amount [161]	excretion concentration | level content decrease | number
kidney [161]	lens eye experiment marrow rat liver bone | infant lense cent
formation [160]	synthesis growth production structure development change rate
system [160]	organ alteration | structure disorder tract presence culture
dose [158]	injection dosage irradiation treatment infusion | kidney
administration [156]	therapy treatment injection | infusion secretion response effect dose
animal [156]	mice | rat dog | rabbit marrow | kidney group result culture patient
data [155]	finding result | experience observation study evidence analysis
diagnosis [154]	etiology finding picture management case evaluation treatment
procedure [154]	method technique surgery | operation criterion treatment
presence [153]	absence relationship reduction increase case concentration effect
infection [152]	characteristic mycoplasma | disease presence type lesion antibody
serum [150]	lense lens | marrow level plasma blood hormone liver pressure value
carcinoma [149]	cancer | tumor | adenocarcinoma disease breast | hyperplasia
finding [148]	data feature examination result observation evidence | diagnosis
fluid [145]	tension | ratio papilloma sinus specimen water concentration
relationship [139]	relation contact | reduction response correlation pattern function
subject [139]	woman child group | man dog patient rabbit lung breast rat
mechanism [136]	finding process factor cause role result | pathogenesis relation
process [136]	mechanism structure change phenomenon | deficiency syndrome
body [135]	structure | line | layer proportion index figure finding type
hypertrophy [135]	hyperplasia | damage insufficiency rise increase dimension retention
infant [130]	child | mice adult fetus rat | age female lense male rabbit
role [128]	part factor | importance pathogenesis relation relationship growth
agent [127]	antibody drug | type experience compound group effect method test
degree [127]	variation reduction rise decrease type period response | amount
problem [126]	disturbance symptom disorder difference form pattern data child
stage [126]	phase period course condition development | tissue week area
material [125]	examination procedure control inclusion layer property type lesion
area [124]	structure population ventricle alteration characteristic | difference
tension [124]	po2 | fluid | output | blood ph vein pressure washing calcium
decrease [121]	increase reduction | rise | fall difference change | value amount
incidence [120]	weight percentage production | risk pathogenesis intensity volume
diabetes [119]	dwarfism fall hr treatment | insufficiency dog hypertrophy
dog [119]	animal | rat mice | subject infant group technique control woman
lung [119]	lens | liver eye serum | nephrectomy culture rat kidney embryo
rise [119]	reduction | increase elevation fall | decrease | retention glucose ratio

MED

word [*Contexts*]	*Groups of closest words*
woman [119]	mother | subject female mice male | year child patient hour rabbit
structure [118]	body surface area | lesion process type pattern layer release system
control [115]	dog | rate material mouse female necrosis child subject patient
lymphocyte [115]	population index percentage transformation | lymphoid nucleus
ratio [115]	concentration rise content | glucose level consumption excretion
examination [113]	analysis finding | procedure study evaluation diagnosis observation
analysis [110]	examination evaluation | observation protein investigation collection
characteristic [109]	difference feature | infection pattern | adult response similarity
condition [109]	prognosis disorder lesion phenomenon feature stage type change
disorder [108]	disturbance abnormality | impairment | alteration function problem
line [108]	type | body ca dash organ tissue proliferation clone amyloidosis
component [107]	property antigen cholesterol source content constituent cause
author [106]	paper case | report fact communication patient infusion hour
hypothermia [106]	brain chemotherapy insufficiency circulation pregnancy information
metabolism [105]	content concentration | mobilization composition synthesis utilization
curve [104]	record | pulse rate gradient | sinus ventricle artery occlusion tension
infusion [103]	route | administration glucose replacement insulin dose artery flow mg
year [103]	hour day month | woman week infant child yr mth | patient
lens [102]	lense | lung eye serum | liver | fraction kidney organ plasma
damage [101]	hypertrophy necrosis | rise utilization increase uptake fibrosis
...	*(skipping to 40's)*
criterion [49]	procedure program | exposure situation dosage efficacy etiology
lipid [49]	cgp | calcium mobilization ca utilization ffa deposition | index
percentage [49]	proportion incidence extent lymphoid estrone | nucleus androsterone
review [49]	report | medulla comparison point | detection survey data application
subtilis [49]	ps | bacterium | organism | culture state dog system | role cell
association [48]	differentiation diagnosis characteristic attempt | variety basis
complication [48]	advantage | duration difficulty month hour consideration obstruction
microfilariae [48]	filariasis interval species | variation individual criterion pool
microscopy [48]	section angiocardiography | sensitivity man | cytoplasm size lung
point [48]	reference | survey review purpose direction finding basis discussion
week [48]	day | yr hr year ffa month | derivative band birth incidence
calcium [47]	ffa ca | phosphate phosphatase osmolality hgh amd | glucose sodium
concept [47]	theory | object program symptom direction idea protocol management
enzyme [47]	mutant migration property component present | substance protein
intake [47]	diet clearance | restriction retention coronary | intensity potassium
nucleus [47]	label percentage cytoplasm | volume equivalent morphology index
rhythm [47]	periodicity cycle | density variety exposure movement gradient arrest
sample [47]	urine reading nickel | output calcium blood present ca biopsy
situation [47]	dwarfs criterion efficacy choice variety prognosis | operation utilization
death [46]	illness | psychosis identification separation life picture environment
hemorrhage [46]	aneurysm illness course abortion replacement | calcium blood hypoxia
medium [46]	ffa ca | requirement mycoplasma day quantity organism concentration
paper [46]	communication | author report experiment | inclusion perfusion
cent [45]	urine age | parathyroidectomy calcium infant parenchyma quarter
exposure [45]	character | irradiation sensitivity uptake damage dosage criterion
field [45]	hemianopia | status palliation subject steroid correction arrest
po2 [45]	tension | oxygen | absorption output | sensitivity dosage utilization
water [45]	total | glucose proportion advantage dosage turnover risk gradient
cortex [44]	label medulla surface | glomerulus capsule diminution index

MED

word [*Contexts*]	*Groups of closest words*
...	*(skipping)*
cytoplasm [37]	surface nucleus | proliferation migration edema comparison label
life [37]	death existence | basis anxiety lung program theory | duration
place [37]	transformation glucose composition | part relation characteristic
wall [37]	endothelium fibrosis | sinus surface | performance myocardium
absorption [36]	exchange po2 calcium | intake na risk | uptake tension transport
bilirubin [36]	thallium | calcium mixture phosphate total oxygen | death relation
combination [36]	onset spray stimulus prognosis | degree presence diagnosis difference
differentiation [36]	association proliferation clone | structure synthesis feature mice
exchange [36]	transport | absorption hg | intake na concentration disappearance
liter [36]	hemophilique | squelette | injection pressure | response rat
metastasis [36]	clone | phenomenon | management carcinoma extract irradiation
occlusion [36]	stenosis constriction obstruction | arrest aorta | hemostasis curve
pathogenesis [36]	incidence tuberculosis survival ca role possibility application
quantity [36]	citrate amount distribution | membrane calcium percentage extent
retardation [36]	decline | phenomenon death sign | ratio weight symptom loss
a-crystallin [35]	antiserum dna line product lack fraction | extract protein animal
cyst [35]	subluxation | history care pressure | loss lung examination year
diet [35]	intake sodium | ion load macrophage saline dash | nickel water
fat [35]	deposition | composition ratio liver population | water glucose
female [35]	male rabbit | woman | infant age lens animal | lense family gland
indication [35]	sign research | measure specificity symptom localization
salt [35]	constituent | ion concentration dosage metabolism | enzyme amount
suspension [35]	inoculation | sarcoma x-irradiate marrow clone jtc-14 culture
tubule [35]	hyperplasia amyloidosis alteration | hypertrophy carcinoma
breast [34]	adenocarcinoma | mice woman subject | extract liver experience
elevation [34]	rise | palliation prolongation gradient | reduction mortality necrosis
need [34]	progress | principle | psychosis concept tension finding management
portion [34]	ca proportion | clone resorption alteration destruction population
program [34]	protocol | way concept criterion experience | year center month
severity [34]	rch course | regard nature evidence | intake feature movement
antiserum [33]	antisera | particle complement antibody | serum placenta lense
attempt [33]	macrophage | progress association record hypothesis | reference
balance [33]	distress record | remission hgh metabolism retention choice
ca [33]	calcium | radioactivity osmolality | ffa mg phosphorus | portion
description [33]	work force | management evaluation report theory palliation evidence
fetus [33]	heart | infant | woman membrane gland lens autism liver | eye adult
ml [33]	egg particle survival | day time component reduction material antigen
record [33]	reading curve balance dilution | scan correlation detection evaluation
regression [33]	remission | improvement bypass | induction necrosis reduction
section [33]	microscopy | comparison angiocardiography passage | line year
status [33]	management field hypertension disappearance regurgitation | series
vitro [33]	insulin interval birth | frequency mg week membrane transformation
assay [32]	fluorescence | radioimmunoassay | molecule inhibition act
detail [32]	rationale illustration | advantage | relation reason consideration
dimension [32]	sinus | angiocardiography performance power | ventricle hypertrophy
division [32]	label proliferation | index hypertrophy | synthesis transport
egg [32]	mycoplasma bacteriophage granuloma crystal | ml mixture nickel
granule [32]	vacuole | equivalent granuloma capacity | membrane nucleus
incorporation [32]	uptake | solution transport species retention | analysis concentration
interaction [32]	communication | person majority find localization performance

MED

word [*Contexts*]	*Groups of closest words*
...	*(skipping)*
hypophysectomy [18]	adrenalectomy | thyroid | mother | infusion experiment growth
injury [18]	anxiety | composition irradiation hypertrophy | symptom structure
left [18]	catheterization | dilatation work obstruction calcium | valve tract
magnitude [18]	detection proportion | absence mechanism reduction site loss
maturation [18]	objective | remission | protection | phage | course lung investigation
medulla [18]	glomerulus | glycoprotein | cortex | region | review hypertrophy
object [18]	concept | cause behavior disorder | child case
pig [18]	state | lymphocyte structure rat | tumor group tissue patient level
placenta [18]	antiserum adult | plasma serum strain | analysis hormone
sulfate [18]	sulphate | ion | loss excretion amount decrease
tolerance [18]	deficiency concentration loss serum | administration activity value
vacuole [18]	granule | inclusion | type number activity fraction cell disease
vessel [18]	size artery flow tension damage | role sign blood fluid plasma
way [18]	arrhythmia | adjunct | program | malformation approach | year
adrenalectomy [17]	hypophysectomy | average | body weight plasma relationship
clone [17]	jtc-14 | metastasis | proliferation suspension portion mice | line
contribution [17]	advance | investigation role rise injection ratio observation
cool [17]	circulation hypothermia | alteration volume | area defect growth
effusion [17]	cavity | tumor patient case cell
environment [17]	death host eye contact | appearance formation therapy factor dna
episode [17]	phenomenon | variation experience stage decrease serum | growth
homograft [17]	kidney correlation | carcinoma dna data child tissue disease
increment [17]	average proportion | composition operation weight hydrocephalus
intensity [17]	calcium | utilization intake titer disappearance dosage incidence
ligation [17]	nephrectomy | pregnancy absence | failure hypertrophy experiment
mass [17]	nucleus | lesion blood | group cell
matter [17]	content case | cell
media [17]	mycoplasma | agent culture | fluid type | method level change
parasite [17]	species | host ventricle origin infection | strain | presence antigen
potassium [17]	ffa calcium | strength intake | ph tension sodium oxygen phosphate
prediction [17]	prognosis | transformation comparison experiment | variation
roentgenogram [17]	technique result tumor study
sheep [17]	ewe | mice | animal lens plasma | rat serum growth content level
skin [17]	laboratory | steroid | brain cent lung bone | injection disorder
supply [17]	cgp | entry | contact removal | production passage population
transmission [17]	onset adult retention toxicity | amyloidosis volume course | rise
urea [17]	serum material amount area | activity patient level cell
worm [17]	species series | strain rate disease group | cell
acidosis [16]	improvement | hypertrophy insufficiency | rise mechanism
acquisition [16]	interpretation | function difference development
cataract [16]	tumor response effect cell
cholesterol [16]	glucose component | present concentration content | acid protein
colitis [16]	case disease change group level
conversion [16]	peak | rise protein characteristic influence damage metabolism
depression [16]	impairment | author agent difference infant body evidence
destruction [16]	deformity | disappearance stasis | calcium ca | interference
education [16]	school | speech | treatment child group
equilibrium [16]	size
evolution [16]	course relation condition species membrane | carcinoma

MED

word [*Contexts*]	*Groups of closest words*
...	*(skipping)*
paralysis [12]	deficiency | response rat day treatment case
physiology [12]	condition | study change response effect cell
prolactin [12]	hormone protein tissue | cell acid
protocol [12]	program | management concept | experience feature surgery
activation [11]	manifestation | tumor case effect
adjunct [11]	way | chemotherapy | infection time therapy procedure
amnion [11]	contact | lung lens relationship culture structure | antigen
angiography [11]	sinus | angiocardiography | catheterization | condition
argument [11]	hypothesis concept | picture | carcinoma observation cancer
chain [11]	content | change
chlorothiazide [11]	thiazide | diuretic | sodium | dog | response type result
column [11]	kidney
delivery [11]	liver | weight volume | level rate patient
deterioration [11]	hydrocephalus rise loss diagnosis | increase change
discrepancy [11]	similarity | support | importance | concept hypothesis
disintegration [11]	disappearance | deficiency | content number formation
distortion [11]	disorder | level development change
diuretic [11]	thiazide | chlorothiazide | administration therapy syndrome
donor [11]	group tumor case cell
drop [11]	count | temperature concentration | change acid level
effort [11]	difference
ego [11]	physical | psychosis | autism death | experience mother
emphasis [11]	reference
energy [11]	anxiety | serum degree blood value concentration | level
fluctuation [11]	variation rise | symptom serum blood difference disease
glycogen [11]	stearate | phospholipid | phosphate amyloidosis | body
hallucination [11]	woman condition form reaction result development
hematoma [11]	dyslexia | area pressure syndrome
hemophilique [11]	squelette | liter
hybrid [11]	suspension | structure decrease | tumor tissue group
hyperbilirubinemia [11]	jaundice
hypercalcemia [11]	picture | infant dog symptom infection formation antigen
hyperparathyroidism [11]	hypoxia | reduction role lesion | defect factor case
idea [11]	reason | concept hypothesis | experience function agent
inscription [11]	growth development type increase change case
insight [11]	experience | data agent pattern treatment therapy study
interrelationship [11]	explanation implication | relation theory nature cause
live [11]	tissue level patient
malignancy [11]	incidence extract serum stage | strain type amount culture
microfilaria [11]	type group case cell patient
mitosis [11]	inhibition relationship lymphocyte synthesis author function
nucleotide [11]	ability
papilloma [11]	aneurysm fluid hyperplasia | report pressure data lesion tumor
pathway [11]	action rate study change
poult [11]	cell patient
propionate [11]	weight
purpura [11]	anemia
reading [11]	record | sample | distribution plasma presence diagnosis
responsiveness [11]	ability nature serum | marrow reaction presence response
sac [11]	sinus | cell sign

MED

word [*Contexts*]	*Groups of closest words*
...	*(skipping)*
influx [9]	transport | production antibody | presence synthesis growth
intermediate [9]	group result level
jtc-14 [9]	clone | suspension existence | transformation | virus characteristic
lead [9]	subject | technique rate group result child effect patient
location [9]	aneurysm | degree case
matrix [9]	analysis protein hormone | increase reaction response
maximum [9]	proportion | variation day drug woman type | dose evidence value
metabolite [9]	excretion
mu [9]	activity
myocardium [9]	performance nephrectomy ventricle consumption | wall
necessity [9]	existence | importance | therapy test increase result activity
one-half [9]	result increase study child patient
organelle [9]	process
outgrowth [9]	line | value response cell
oxidation [9]	phosphorylation | availability | synthesis concentration level
pad [9]	tissue strain tumor cell
plan [9]	role serum method form period rate group | result effect patient
poison [9]	foci | failure | infection group
polymerase [9]	dna | content | effect
polyuria [9]	dwarfism | relation management | hydrocephalus appearance
precipitation [9]	acid
psychotherapy [9]	death association | experience mother culture value infection
radioimmunoassay [9]	assay | unit analysis | hypothermia technique test procedure
reactivity [9]	relationship volume | syndrome rate activity response
read [9]	response
rejection [9]	reaction growth
resonance [9]	pattern
ribosome [9]	nickel | ratio | activity cell
saline [9]	urine | consumption | diet | age plasma condition dose | test
sense [9]	part | observation | group patient
sequelae [9]	characteristic pattern result | rate change patient cell
serotonin [9]	hormone | growth effect change
sibling [9]	family absence | subject syndrome evidence development | child
spectrum [9]	antigen tumor reaction response study cell
spite [9]	study patient
subluxation [9]	cyst | symptom lesion | disease case
t-loop [9]	ordination | absence | case effect
tetralogy [9]	case
trace [9]	preparation | virus loss | difference change
trend [9]	increase
ultrafiltrate [9]	method cell
vitamin [9]	hypothermia agent pattern procedure hormone dna | acid
x-irradiate [9]	migration suspension | range mice | age virus epithelium
abortion [8]	hemorrhage | effect
ahf [8]	type lesion tumor | patient
allergy [8]	tissue
amino [8]	subject
antibiotic [8]	agent | change
arteritis [8]	structure formation defect | lesion reaction disease
article [8]	animal subject

MED

word [*Contexts*]	*Groups of closest words*
...	*(skipping)*
stroma [5]	transplantation | removal | growth treatment response
t-960 [5]	organism mycoplasma | infection
technics [5]	technique time test level
thallium [5]	bilirubin | incidence weight body flow | time rate change
threat [5]	treatment
discontinuance [4]	therapy patient
documentation [4]	patient
electroencephalogram [4]	finding
ewe [4]	sheep
feed [4]	level
filter [4]	tumor
glomerulus [4]	medulla | cortex | hypertrophy | action growth result
glycolysis [4]	production | growth
guide [4]	method
haematein [4]	distribution analysis
hs [4]	carcinoma antigen | cell
hypercapnia [4]	effect
hyperglycemia [4]	sign | decrease form growth | test increase case change
igg [4]	blood
ii-deoxy-17-oxosteroid [4]	level
immunoelectrophoresis [4]	formation method tumor result group case change
immunofluorescence [4]	effect
inference [4]	study
ligand [4]	expression
london [4]	pattern rate change | level
long-term [4]	effect
lsd-25 [4]	distribution | response disease case change child effect
malizia [4]	cell
necropsy [4]	cancer
neurolyticum [4]	characteristic presence | effect
neutrophil [4]	cell
nigra [4]	type case
oleic [4]	acid
operant [4]	report | case
ordination [4]	t-loop | disease case
osmiophilic [4]	type
parasitaemia [4]	change
phocyte [4]	hormone cell
purity [4]	composition
raise [4]	lower
rationale [4]	detail | finding result
recrudescence [4]	serum administration test increase | change study
room [4]	rate
serotype [4]	strain
shrinkage [4]	response change
stearate [4]	glycogen | phospholipid | capacity body | activity
stem [4]	function
twothird [4]	nucleus | characteristic | change cell

4

SEMANTIC CLUSTERING

SEMANTIC CLUSTERING OVER THE MED CORPUS	
Semantic Axis	*words closest to axis*
a-crystallin *as a* dna	protein
ability *as a* capacity	production function
ability *as a* inability	capacity
abnormality *as a* anomaly	atresia
abnormality *as a* impairment	disorder disturbance
abnormality *as a* nature	manifestation
absence *as a* sibling	family
absorption *as a* exchange	transport
absorption *as a* na	exchange
absorption *as a* po	tension
accumulation *as a* extent	jaundice
acid *as a* dna	protein
acid *as a* fraction	protein
acidosis *as a* insufficiency	hypertrophy
act *as a* prolongation	deficiency
activity *as a* amount	concentration level number
addition *as a* absence	presence
adenocarcinoma *as a* carcinoma	tumor
adenoma *as a* hyperplasia	hypertrophy
adjunct *as a* chemotherapy	therapy
administration *as a* dose	injection
administration *as a* infusion	dose
administration *as a* secretion	deficiency
administration *as a* therapy	treatment
administration *as a* treatment	response
affection *as a* psychosis	disorder
aggregation *as a* present	ability
agreement *as a* discussion	interpretation basis
agreement *as a* interpretation	significance discussion basis
agreement *as a* significance	correlation
antiserum *as a* lense	serum
anxiety *as a* separation	experience
application *as a* advance	experience
application *as a* importance	feature
approach *as a* palliation	management chemotherapy
atresia *as a* amputation	duct dystrophy
atresia *as a* aneurysm	obstruction regurgitation
atresia *as a* dystrophy	amputation family
atresia *as a* malformation	valve regurgitation
atresia *as a* obstruction	valve aneurysm
atresia *as a* valve	obstruction malformation

Semantic Axis	*words closest to axis*
author *as a* communication	paper report
author *as a* paper	report communication experiment
author *as a* report	experiment
autism *as a* childhood	psychosis schizophrenia
autism *as a* ego	psychosis
autism *as a* psychosis	schizophrenia childhood
autism *as a* schizophrenia	psychosis childhood
availability *as a* utilization	uptake metabolism
average *as a* intensity	utilization intake
average *as a* utilization	intensity ffa
bacteriophage *as a* phage	strain
bacteriophage *as a* ps	phage
bacterium *as a* organism	subtilis
bacterium *as a* subtilis	organism
balance *as a* hgh	calcium
balance *as a* record	measurement
band *as a* filtration	disappearance
basis *as a* discussion	interpretation
basis *as a* interpretation	discussion
basis *as a* regard	relation reference
behavior *as a* disorder	function
biopsy *as a* parenchyma	transplantation
biopsy *as a* transplantation	parenchyma
blood *as a* liver	serum plasma
blood *as a* plasma	serum liver
blood *as a* serum	level concentration
blood *as a* tension	pressure
blood *as a* volume	pressure
cause *as a* etiology	nature diagnosis
cause *as a* nature	abnormality
cavity *as a* chamber	eye
cavity *as a* eye	chamber heart lung lens kidney
cavity *as a* organ	lens
cell *as a* type	case
cent *as a* age	infant
cent *as a* lense	infant kidney
center *as a* attempt	hypothesis
chamber *as a* cavity	eye
chamber *as a* eye	heart
chamber *as a* series	heart
change *as a* decrease	increase difference rise
change *as a* difference	increase
change *as a* increase	effect response
change *as a* pattern	difference
change *as a* response	increase study effect
change *as a* study	effect
characteristic *as a* decrease	difference change
characteristic *as a* difference	change
characteristic *as a* feature	pattern
characteristic *as a* pattern	difference change

Semantic Axis	*words closest to axis*
chemotherapy *as a* radiotherapy	palliation
chemotherapy *as a* route	infusion
child *as a* case	patient
child *as a* group	patient result
child *as a* subject	group
child *as a* woman	patient subject
child *as a* year	woman
childhood *as a* psychosis	schizophrenia autism syndrome
childhood *as a* schizophrenia	psychosis autism
chlorothiazide *as a* diuretic	thiazide
chlorothiazide *as a* thiazide	diuretic
choice *as a* majority	selection
choice *as a* prognosis	evaluation management
choice *as a* selection	majority
chromosome *as a* titer	turnover incidence
chromosome *as a* turnover	titer
circulation *as a* cool	hypothermia
cirrhosis *as a* neoplasm	amyloidosis
citrate *as a* output	volume
clearance *as a* excretion	concentration
clearance *as a* permeability	reabsorption
clearance *as a* ratio	excretion concentration content
clearance *as a* reabsorption	excretion
clone *as a* differentiation	proliferation
clone *as a* jtc-	suspension
clone *as a* line	culture
clone *as a* proliferation	differentiation
clone *as a* suspension	culture
closure *as a* malformation	defect
communication *as a* author	report
communication *as a* paper	report author experiment
communication *as a* report	author experiment data
comparison *as a* survey	review
complication *as a* advantage	consideration
component *as a* antigen	fraction
component *as a* constituent	property
component *as a* fraction	antigen content
component *as a* synthesis	content
composition *as a* distribution	property
composition *as a* metabolism	content
composition *as a* mobilization	metabolism
concentration *as a* amount	level excretion
concentration *as a* content	level
concentration *as a* excretion	level
concentration *as a* metabolism	content synthesis
concentration *as a* rate	level
concentration *as a* ratio	content excretion
concentration *as a* serum	level
concentration *as a* synthesis	content
concentration *as a* value	level rate

Semantic Axis	*words closest to axis*
concept *as a* protocol	program management
conclusion *as a* diagnosis	finding
consideration *as a* advantage	detail
consideration *as a* detail	advantage
consideration *as a* explanation	significance implication
consideration *as a* implication	explanation
constituent *as a* component	antigen fraction
constituent *as a* property	component
consumption *as a* dimension	power
consumption *as a* myocardium	wall
consumption *as a* power	dimension
consumption *as a* utilization	uptake
contact *as a* relation	relationship
content *as a* composition	metabolism
content *as a* concentration	level
content *as a* fraction	protein acid
content *as a* metabolism	concentration synthesis
content *as a* protein	acid
content *as a* ratio	concentration weight
content *as a* synthesis	concentration
control *as a* dog	animal subject
control *as a* group	child
cool *as a* circulation	hypothermia
coronary *as a* flow	volume pressure
correction *as a* localization	field
correlation *as a* reduction	relationship rise
cortex *as a* capsule	surface
cortex *as a* glomerulus	medulla
cortex *as a* hyperplasia	hypertrophy
cortex *as a* label	index
cortex *as a* medulla	glomerulus
course *as a* aneurysm	hemorrhage
course *as a* hemorrhage	aneurysm
course *as a* time	day
criterion *as a* efficacy	situation
crystal *as a* egg	mycoplasma
culture *as a* suspension	marrow
curve *as a* pressure	rate
curve *as a* record	dilution
curve *as a* ventricle	artery
cytoplasm *as a* edema	surface migration
cytoplasm *as a* proliferation	migration label
dash *as a* line	culture
data *as a* analysis	observation technique
data *as a* finding	result observation evidence
data *as a* observation	result study
data *as a* result	study
day *as a* group	patient
day *as a* hour	year
day *as a* hr	hour month week
day *as a* month	hour week year time

Semantic Axis	*words closest to axis*
day *as a* week	hr month year
day *as a* year	hour
day *as a* yr	hr hour month week year
decrease *as a* amount	concentration
decrease *as a* characteristic	difference
decrease *as a* difference	increase change
decrease *as a* fall	reduction rise
decrease *as a* increase	change
decrease *as a* reduction	increase rise
decrease *as a* rise	increase difference
decrease *as a* value	increase concentration
defect *as a* hemianopia	field
defect *as a* malformation	regurgitation
defect *as a* type	case
deficiency *as a* secretion	administration
deficit *as a* impairment	disturbance
degree *as a* variation	reduction
demonstration *as a* contact	range
demonstration *as a* determination	measurement
density *as a* ph	flow
density *as a* survival	weight
dependence *as a* correlation	difference
destruction *as a* interference	portion
destruction *as a* portion	ca interference
detail *as a* advantage	reason consideration
detail *as a* illustration	advantage
detail *as a* reason	advantage
detection *as a* dimension	power volume
detection *as a* power	dimension
detection *as a* record	measurement
determination *as a* estimation	measurement
determination *as a* measurement	analysis
development *as a* production	incidence
diagnosis *as a* etiology	picture prognosis
diagnosis *as a* management	prognosis
diagnosis *as a* picture	etiology
diagnosis *as a* prognosis	etiology management evaluation
diagnosis *as a* treatment	case
diagnostic *as a* surgery	procedure
diet *as a* intake	sodium clearance
diet *as a* load	nickel
diet *as a* sodium	intake
difference *as a* characteristic	pattern decrease
difference *as a* decrease	change characteristic increase rise
difference *as a* increase	change
difference *as a* observation	pattern
difference *as a* pattern	change observation
difference *as a* rise	increase decrease
differentiation *as a* clone	proliferation

Semantic Axis	*words closest to axis*
dilatation *as a* aneurysm	obstruction regurgitation
dilatation *as a* insufficiency	regurgitation
dilatation *as a* left	catheterization obstruction tract
dilatation *as a* obstruction	tract aneurysm
dilatation *as a* tract	obstruction
dimension *as a* angiocardiography	sinus performance
dimension *as a* detection	volume
dimension *as a* performance	sinus angiocardiography power
dimension *as a* power	performance detection consumption
dimension *as a* sinus	angiocardiography performance
discrepancy *as a* similarity	characteristic
discrepancy *as a* support	hypothesis
discussion *as a* agreement	significance interpretation experience
discussion *as a* interpretation	basis significance
disease *as a* case	patient
disease *as a* lesion	case
disease *as a* type	case
disorder *as a* alteration	function
disorder *as a* behavior	function
disorder *as a* disturbance	problem
disorder *as a* impairment	disturbance abnormality
disorder *as a* psychosis	disturbance
distribution *as a* composition	content property
disturbance *as a* deficit	impairment
disturbance *as a* disorder	problem alteration
disturbance *as a* feature	pattern
disturbance *as a* impairment	disorder
disturbance *as a* psychosis	disorder
diuretic *as a* chlorothiazide	thiazide
diuretic *as a* thiazide	chlorothiazide
division *as a* index	nucleus
division *as a* label	proliferation index nucleus
division *as a* proliferation	label
dna *as a* antigen	protein hormone fraction
dna *as a* molecule	synthesis
dna *as a* protein	hormone acid
dog *as a* animal	rat
dog *as a* control	group
dog *as a* mice	animal infant woman mouse
dog *as a* mouse	rat mice
dog *as a* subject	group
dosage *as a* intensity	titer
dosage *as a* titer	intensity
dose *as a* administration	injection treatment
dose *as a* infusion	administration
dose *as a* radiation	irradiation
drug *as a* chemotherapy	infusion hypothermia
duct *as a* amputation	atresia management
duct *as a* family	atresia

Semantic Axis	*words closest to axis*
duration *as a* survival	retention
dwarf *as a* male	mice mouse infant
dwarfs *as a* secretion	deficiency
dystrophy *as a* amputation	management atresia
dystrophy *as a* family	atresia prognosis
dystrophy *as a* management	prognosis
edema *as a* cytoplasm	surface
edema *as a* inclusion	migration
edema *as a* lining	surface hyperplasia
edema *as a* migration	inclusion
edema *as a* surface	lining
effect *as a* change	study
effect *as a* increase	change response level
effect *as a* response	change result increase study
effect *as a* result	study
effect *as a* value	response level increase
efficacy *as a* criterion	situation
efficacy *as a* situation	criterion
efficiency *as a* risk	average
egg *as a* mycoplasma	infection
ego *as a* psychosis	autism death
element *as a* fact	author
elevation *as a* mortality	necrosis
elevation *as a* necrosis	mortality
embryo *as a* brain	bone
embryo *as a* epithelium	lens
embryo *as a* sarcoma	age
embryo *as a* spleen	bone
entry *as a* label	molecule
entry *as a* requirement	deficiency
enzyme *as a* property	component
epithelium *as a* eye	lens lung lense
epithelium *as a* lens	lung
estimation *as a* determination	measurement
etiology *as a* nature	cause
etiology *as a* picture	diagnosis
etiology *as a* prognosis	diagnosis
evaluation *as a* analysis	observation
evaluation *as a* investigation	analysis
evaluation *as a* measurement	analysis
evaluation *as a* prognosis	diagnosis
evaluation *as a* record	measurement
evidence *as a* data	finding result
evidence *as a* feature	finding
evidence *as a* finding	data result
examination *as a* analysis	observation
examination *as a* finding	diagnosis observation
examination *as a* record	evaluation
exception *as a* explanation	relation
exchange *as a* absorption	po

5

AUTOMATIC THESAURUS GENERATION

Thesaurus Generated from the MED Corpus
First 150 Most Frequent Words

acid :: [486 contexts, frequency rank 8] MED *Relat.* dna, fraction, hormone, activity, protein. *Vbs.* saturate, transform, mobilize, increase, extract, esterify. *Exp.* amino acid (cf. testosterone propionate, factor viii), tenuazonic acid (cf. tumor growth, vit d), acid phosphatase (cf. enzyme activity, electron microscopy), acid metabolism (cf. mean concentration, folic acid), folic acid (cf. rat kidney, dna content), acid composition (cf. total lipid, blood glucose).

action :: [166 contexts, frequency rank 57] MED *Relat.* effect; influence, ability. *Exp.* action potential (cf. time constant, coronary flow)

activity :: [410 contexts, frequency rank 11] MED *Relat.* level, effect; protein, concentration, amount, number. *Vbs.* increase, show, determine, decrease, reduce, inhibit, enhance, contain, alter. *Exp.* enzyme activity (cf. acid phosphatase, testosterone propionate), surface activity (cf. surface tension, inclusion body).

administration :: [156 contexts, frequency rank 62] MED *Relat.* dose; injection, response, treatment, therapy; deficiency, secretion, infusion. *Vbs.* follow, associate.

agent :: [127 contexts, frequency rank 76] MED *Relat.* compound, drug, antibody. *Vbs.* produce, use.

alteration :: [99 contexts, frequency rank 97] MED *Relat.* deficiency, disorder, reduction; function, change, increase, rise; . *Vbs.* observe.

amount :: [161 contexts, frequency rank 59] MED *Relat.* number, excretion; activity, increase, level, concentration; correlation, decrease. *Vbs.* increase, contain, excrete.

analysis :: [110 contexts, frequency rank 86] MED *Relat.* examination; data, test, technique, protein, observation; collection, measurement, investigation, evaluation. *Vbs.* make, reveal.

animal :: [156 contexts, frequency rank 62] MED *Relat.* culture, rat; rabbit, dog, mice. *Vbs.* treat, irradiate, infect.

antibody :: [95 contexts, frequency rank 101] MED *Relat.* part; agent, reaction, antigen; antisera, inhibitor, antiserum. *Vbs.* demonstrate. *Fam.* antigen, antiserum.

antigen :: [184 contexts, frequency rank 48] MED *Relat.* culture, fraction, dna, protein; growth, hormone, reaction; virus, component, antibody. *Vbs.* demonstrate, test, represent, react, contain. *Fam.* antibody, antigenic.

area :: [124 contexts, frequency rank 79] MED *Relat.* stage, structure; part, population.

artery :: [96 contexts, frequency rank 100] MED *Relat.* hyperplasia; flow; aorta, nerve, hypoplasia, superior, ventricle, chamber, vein. *Exp.* coronary artery (cf. blood flow, total flow)

author :: [106 contexts, frequency rank 90] MED *Relat.* experiment; fact, communication, report, paper. *Vbs.* believe, use, study, discuss, describe.

basis :: [85 contexts, frequency rank 109] MED *Relat.* relation; reference, regard, interpretation, discussion. *Vbs.* discuss, calculate.

blood :: [258 contexts, frequency rank 27] MED *Relat.* level; liver, plasma, marrow, value, serum, oxygen, tension. *Vbs.* increase, study, make, find, estimate. *Exp.* blood pressure (cf. oxygen tension, carbon dioxide), blood flow (cf. carbon dioxide, fluid po2), blood volume (cf. stroke volume, blood glucose), blood glucose (cf. newborn lamb, acid composition), peripheral blood (cf. bone marrow, type ii), cord blood (cf. total lipid, protein metabolism), blood pool (cf. blood volume, cell population), blood viscosity (cf. fluid pressure, oxygen consumption), blood stream (cf. electron microscope, plasma cell).

body :: [135 contexts, frequency rank 73] MED *Relat.* type; layer, line, structure. *Vbs.* contain. *Exp.* inclusion body (cf. surface activity, type ii), body weight (cf. kidney weight, dna content), body image (cf. visual agnosia, separation anxiety), body temperature (cf. extracorporeal circulation, flow rate), body growth (cf. tumor growth, body weight).

cancer :: [255 contexts, frequency rank 29] MED *Relat.* lesion, tumor; tissue, disease; carcinoma. *Vbs.* advance, disseminate. *Exp.* cancer patient (cf. survival time, joint deformity), cancer chemotherapy (cf. survival time, intra-arterial infusion), cancer cell (cf. human cell, year period).

carcinoma :: [149 contexts, frequency rank 68] MED *Relat.* disease, tissue, lesion, tumor, cancer; hyperplasia, breast, adenocarcinoma. *Vbs.* advance. *Exp.* cell carcinoma (cf. cell line, human lung)

case :: [572 contexts, frequency rank 5] MED *Relat.* change, study; patient; result, treatment, child, defect, type, disease, lesion. *Vbs.* present, report, occur, find, describe, study, discuss, use, observe, classify, diagnose, analyze. *Exp.* case report (cf. intra-arterial infusion, age group), case history (cf. compound lipid, inclusion disease), index case (cf. cleft palate, childhood schizophrenia).

cell :: [1156 contexts, frequency rank 1] MED *Relat.* tissue. *Vbs.* label, find, infect, contain, appear, show, nucleate, culture, transfuse, transform, observe, make. *Exp.* liver cell (cf. adult patient, bone resorption), cell line (cf. cell carcinoma, tissue culture), hela cell (cf. human lung, lymph node), cell culture (cf. mycoplasma strain, actinomycin d), cell division (cf. zona glomerulosa, folic acid), spleen cell (cf. bone resorption, liver cell), cell type (cf. vit d, survival rate), mast cell (cf. surface tension, inclusion body), plasma cell (cf. blood stream, surface activity), human cell (cf. lung tissue, human lung).

change :: [549 contexts, frequency rank 6] MED *Relat.* study, effect; alteration, disease, pattern, rise, decrease, difference, response, increase. *Vbs.* occur, observe, show, produce, find, result, mark, induce, associate, reveal, relate, note.

characteristic :: [109 contexts, frequency rank 87] MED *Relat.* decrease; infection, type, pattern, difference; course, similarity, adult, feature. *Vbs.* induce.

child :: [412 contexts, frequency rank 10] MED *Relat.* result, group; case, patient; reaction, year, woman, form, subject, infant. *Vbs.* disturb, show, study, observe, give, bear, report, present, match, find, diagnose, develop. *Fam.* childhood.

component :: [107 contexts, frequency rank 89] MED *Relat.* content, synthesis, fraction, antigen; cause, constituent, source, cholesterol, property. *Vbs.* contain, consist. *Exp.* protein component (cf. wuchereria bancrofti, skin reaction)

concentration :: [339 contexts, frequency rank 20] MED *Relat.* rate; level; amount, synthesis, metabolism, rise, ratio, value, excretion, content. *Vbs.* increase, decrease, lower, present, find, contain. *Exp.* sm concentration (cf. ffa level, phage dna), plasma concentration (cf.

urine volume, sodium intake), mean concentration (cf. acid metabolism, acid composition), dna concentration (cf. phage dna, dna content).

condition :: [109 contexts, frequency rank 87] MED *Relat.* stage, disorder; phenomenon, prognosis. *Vbs.* use, develop.

content :: [212 contexts, frequency rank 36] MED *Relat.* fraction; level, concentration; glucose, rise, weight, ratio, composition, synthesis, metabolism. *Vbs.* increase. *Exp.* dna content (cf. growth retardation, folic acid), total content (cf. dna content, kidney weight).

control :: [115 contexts, frequency rank 84] MED *Relat.* material, dog; subject; . *Vbs.* compare, serve. *Exp.* control group (cf. total estrogen, vit d), control kidney (cf. dna content, rat kidney).

correlation :: [93 contexts, frequency rank 103] MED *Relat.* reduction; amount, rise, relationship, difference; adjustment, dependence, survey, record, significance. *Vbs.* show, find.

course :: [93 contexts, frequency rank 103] MED *Relat.* characteristic; stage; recovery, aneurysm, hemorrhage, abnormality, severity.

culture :: [208 contexts, frequency rank 37] MED *Relat.* marrow; animal, specimen, antigen, lung, extract, suspension. *Vbs.* infect, isolate. *Exp.* tissue culture (cf. human lung, electron microscopy), cell culture (cf. mycoplasma strain, actinomycin d).

curve :: [104 contexts, frequency rank 92] MED *Relat.* artery; nomogram, ventricle, gradient, pulse, record. *Vbs.* obtain. *Exp.* pressure curve (cf. right ventricle, left ventricle), dilution curve (cf. right ventricle, left ventricle).

damage :: [101 contexts, frequency rank 95] MED *Relat.* hypertrophy; uvr, infiltration, fibrosis, uptake, necrosis. *Vbs.* induce, result. *Exp.* brain damage (cf. childhood schizophrenia, heart rate)

data :: [155 contexts, frequency rank 63] MED *Relat.* evidence, observation, finding; technique, study, result; problem, report, analysis, experience. *Vbs.* obtain, suggest, indicate, present.

day :: [203 contexts, frequency rank 41] MED *Relat.* time; rat, group, patient; yr, year, week, month, hour, hr. *Vbs.* return, follow, occur, reach, maintain, find, carry.

decrease :: [121 contexts, frequency rank 80] MED *Relat.* characteristic, rise; amount, value, concentration, difference, change, increase; fall, reduction. *Vbs.* show, accompany.

defect :: [338 contexts, frequency rank 21] MED *Relat.* case; anomaly, type, regurgitation, malformation, disorder, disturbance. *Vbs.* isolate, associate, mark.

deficiency :: [99 contexts, frequency rank 97] MED *Relat.* measurement, alteration, loss; administration; requirement, ahf, mode, secretion. *Vbs.* isolate, associate. *Fam.* deficient.

degree :: [127 contexts, frequency rank 76] MED *Relat.* period; reduction, variation. *Vbs.* vary, show.

development :: [248 contexts, frequency rank 31] MED *Relat.* growth; stage, incidence, production. *Exp.* language development (cf. separation anxiety, childhood schizophrenia)

diabetes :: [119 contexts, frequency rank 82] MED *Relat.* intramuscular, insoluble, difficulty, dwarfism, diuretic, thiazide, chlorothiazide.

diagnosis :: [154 contexts, frequency rank 64] MED *Relat.* finding; case, treatment; conclusion, significance, evaluation, prognosis, management, picture, etiology. *Vbs.* make, suggest, confirm. *Fam.* diagnostic.

difference :: [216 contexts, frequency rank 35] MED *Relat.* group, increase, change; variation, observation, correlation, rise, decrease, pattern, characteristic. *Vbs.* find, reveal, observe, note, mark. *Fam.* different.

disease :: [401 contexts, frequency rank 12] MED *Relat.* change, patient, case; infection, tissue, type, carcinoma, lesion. *Vbs.* suffer, report, relate, produce, affect. *Exp.* heart disease (cf. nervous system, outflow tract), inclusion disease (cf. newborn infant, steroid therapy), collagen disease (cf. testosterone propionate, inclusion body).

disorder :: [108 contexts, frequency rank 88] MED *Relat.* condition, alteration, problem; function, defect; disability, psychosis, impairment, abnormality, disturbance.

distribution :: [70 contexts, frequency rank 124] MED *Relat.* composition; loss, property, content; toxicosis, pathology.

disturbance :: [80 contexts, frequency rank 114] MED *Relat.* pattern, feature, alteration, problem, defect, disorder; deficit, aspect, impairment.

dna :: [193 contexts, frequency rank 45] MED *Relat.* fraction, antigen, protein; acid, hormone; polymerase. *Vbs.* transform, synthesize, isolate. *Exp.* dna content (cf. growth retardation, folic acid), phage dna (cf. dna concentration, dna molecule), dna concentration (cf. phage dna, dna content), dna molecule (cf. phage dna, protein metabolism), dna label (cf. cell population, lymph node).

dog :: [119 contexts, frequency rank 82] MED *Relat.* control, subject, woman, infant; technique, group, rat, animal; mouse, mice. *Vbs.* irradiate, survive, cool, study, infect. *Exp.* donor dog (cf. extracorporeal circulation, nickel carbonyl)

dose :: [158 contexts, frequency rank 61] MED *Relat.* weight, administration; therapy, treatment, injection; infusion, irradiation. *Vbs.* give, increase, tolerate, receive, absorb.

effect :: [650 contexts, frequency rank 3] MED *Relat.* study; value, growth, activity, action, level, increase, result, response, change. *Vbs.* produce, exert, determine, study, note, make, suggest, prevent, enhance, demonstrate, abolish. *Exp.* side effect (cf. initial value, bile duct)

evaluation :: [67 contexts, frequency rank 126] MED *Relat.* investigation; measurement, diagnosis, observation, analysis; range, palliation, world, prognosis.

evidence :: [177 contexts, frequency rank 53] MED *Relat.* result; cause, severity, information, data, finding, feature. *Vbs.* present, show, reveal, provide, take, suggest, give, find.

examination :: [113 contexts, frequency rank 85] MED *Relat.* material, analysis; procedure, diagnosis, finding, observation; manifestation. *Vbs.* show, reveal, carry.

excretion :: [184 contexts, frequency rank 48] MED *Relat.* volume; level, concentration; production, retention, reabsorption, ratio, clearance, amount. *Vbs.* increase, reduce. *Exp.* protein excretion (cf. adult patient, filtration rate) *Fam.* excrete.

experiment :: [96 contexts, frequency rank 100] MED *Relat.* report; marrow, kidney; theory, paper. *Vbs.* perform, suggest. *Fam.* experimental.

extract :: [98 contexts, frequency rank 98] MED *Relat.* preparation; hormone, culture; gland, homogenate, property. *Exp.* parathyroid extract (cf. parathyroid hormone, actinomycin d)

factor :: [236 contexts, frequency rank 33] MED *Relat.* aspect, mechanism, role. *Vbs.* influence, relate, involve, identify. *Exp.* factor viii (cf. amino acid, growth hormone)

feature :: [97 contexts, frequency rank 99] MED *Relat.* characteristic; type, evidence, finding, pattern; disturbance, abnormality, importance, picture, grade.

finding :: [148 contexts, frequency rank 69] MED *Relat.* diagnosis, evidence, data; observation, result; appearance, sign, examination, feature. *Vbs.* discuss, suggest, support, present, observe.

flow :: [177 contexts, frequency rank 53] MED *Relat.* volume; pressure; oxygen, artery, circulation, ph. *Vbs.* increase, reduce. *Exp.* blood flow (cf. carbon dioxide, fluid po2), flow rate (cf. coronary flow, body temperature), total flow (cf. coronary artery, left ventricle), coronary flow (cf. oxygen consumption, flow rate).

fluid :: [145 contexts, frequency rank 70] MED *Relat.* blood; water, papilloma, tension. *Vbs.* shunt, differ, contain. *Exp.* fluid pressure (cf. carbon dioxide, blood flow), fluid po2 (cf. carbon dioxide, blood flow).

form :: [198 contexts, frequency rank 43] MED *Relat.* patient, child, disease, case, type; problem, finding, sign. *Vbs.* take, reflect, observe, grow. *Fam.* formation.

formation :: [160 contexts, frequency rank 60] MED *Relat.* synthesis; . *Fam.* form.

fraction :: [196 contexts, frequency rank 44] MED *Relat.* antigen, content; acid, concentration; ability, component, preparation, serum, lens. *Vbs.* show, separate, measure, hydrolyze. *Exp.* rna fraction (cf. body growth, total content), protein fraction (cf. insoluble protein, ionic strength).

function :: [176 contexts, frequency rank 54] MED *Relat.* response; behavior, stem, relationship, alteration, disorder, ability, capacity. *Vbs.* preserve. *Fam.* functional.

group :: [397 contexts, frequency rank 13] MED *Relat.* result, child; case, patient; type, day, subject, difference. *Vbs.* show, find, divide, compare, classify, use, select, occur, obtain, follow. *Exp.* age group (cf. cell type, time constant), control group (cf. total estrogen, vit d).

growth :: [284 contexts, frequency rank 25] MED *Relat.* tumor; effect, tissue; antigen, protein, development. *Vbs.* retard, stimulate, show, follow, enhance, accelerate. *Exp.* growth hormone (cf. bone marrow, parathyroid hormone), growth rate (cf. growth retardation, folic acid), tumor growth (cf. body growth, tenuazonic acid), growth retardation (cf. dna content, body weight), body growth (cf. tumor growth, body weight).

hormone :: [365 contexts, frequency rank 17] MED *Relat.* extract, dna, antigen, protein, serum. *Vbs.* label, administer, produce, increase, contain. *Exp.* growth hormone (cf. bone marrow, parathyroid hormone), parathyroid hormone (cf. plasma calcium, vitamin d), steroid hormone (cf. compound lipid, control kidney), hormone therapy (cf. intra-arterial infusion, steroid therapy).

hour :: [90 contexts, frequency rank 105] MED *Relat.* year; woman, day; find, nephrectomy, yr, week, month, hr. *Vbs.* occur, appear.

hydrocephalus :: [84 contexts, frequency rank 110] MED *Relat.* heart, hepatitis, jaundice. *Vbs.* develop. *Fam.* hydrocephalic.

hyperplasia :: [90 contexts, frequency rank 105] MED *Relat.* artery; carcinoma, hypertrophy; origin, circulation, adenoma, amyloidosis.

hypertrophy :: [135 contexts, frequency rank 73] MED *Relat.* enlargement, rise, retention, dimension, insufficiency, damage, hyperplasia. *Vbs.* show.

hypothermia :: [106 contexts, frequency rank 90] MED *Relat.* therapy; perfusion, circulation, cool, information, chemotherapy, brain. *Vbs.* use. *Fam.* hypothermic.

incidence :: [120 contexts, frequency rank 81] MED *Relat.* volume, weight; pathogenesis, risk, production, percentage. *Vbs.* reduce, increase.

increase :: [385 contexts, frequency rank 16] MED *Relat.* rate, response; level, effect, change; reduction, value, difference, rise, decrease. *Vbs.* show, cause, result, mark, indicate, find, take, produce, occur, observe, lead, induce.

infant :: [130 contexts, frequency rank 74] MED *Relat.* rat, child; lense, rabbit, male, female, age, fetus, adult, mice. *Vbs.* treat, operate. *Exp.* newborn infant (cf. inclusion disease, type ii) *Fam.* infancy.

infection :: [152 contexts, frequency rank 66] MED *Relat.* case, disease; mycoplasma, characteristic. *Vbs.* cause. *Fam.* infect.

infusion :: [103 contexts, frequency rank 93] MED *Relat.* dose, administration; irradiation, mg, replacement, route. *Vbs.* prolong, use. *Exp.* intra-arterial infusion (cf. cancer chemotherapy, hormone therapy)

injection :: [205 contexts, frequency rank 39] MED *Relat.* time; number, dose, administration. *Vbs.* follow, receive, give, make.

investigation :: [74 contexts, frequency rank 120] MED *Relat.* evaluation; mechanism, observation, analysis; knowledge.

irradiation :: [77 contexts, frequency rank 117] MED *Relat.* infusion, dose; management, exposure, x ray, x-irradiation, radiation. *Vbs.* receive, follow. *Fam.* irradiate.

kidney :: [161 contexts, frequency rank 59] MED *Relat.* marrow; rat; cent, infant, experiment, lense, bone, eye, liver, lens. *Vbs.* increase, find, compensate. *Exp.* kidney weight (cf. dna content, body weight), rat kidney (cf. folic acid, testosterone propionate), kidney cell (cf. human cell, dna content), control kidney (cf. dna content, rat kidney).

lens :: [102 contexts, frequency rank 94] MED *Relat.* plasma, liver, lung; fraction, kidney, serum; regeneration, organ, eye, lense. *Exp.* lens protein (cf. acid metabolism, lens epithelium), lens regeneration (cf. electron microscope, protein fraction), lens epithelium (cf. lymph node, folic acid). *Fam.* lense.

lesion :: [258 contexts, frequency rank 27] MED *Relat.* tumor, cancer; study, change, case, disease; manifestation, symptom. *Vbs.* find, result, develop.

level :: [548 contexts, frequency rank 7] MED *Relat.* effect; serum, amount, blood, content, excretion, increase, value, rate, concentration. *Vbs.* increase, reduce, excrete, use, reach, determine, decrease, result, elevate, achieve, vary, produce. *Exp.* ffa level (cf. blood glucose, control group)

line :: [108 contexts, frequency rank 88] MED *Relat.* body, type; organ, dash. *Vbs.* establish, show, obtain. *Exp.* cell line (cf. cell carcinoma, tissue culture)

liver :: [100 contexts, frequency rank 96] MED *Relat.* plasma, lung, lens; tissue, blood, marrow, kidney, serum; age, spleen. *Vbs.* increase. *Exp.* liver cell (cf. adult patient, bone resorption), liver biopsy (cf. inclusion disease, lung tissue).

loss :: [98 contexts, frequency rank 98] MED *Relat.* deficiency; rise; distribution, fall, suppression, retention. *Vbs.* hear.

lung :: [119 contexts, frequency rank 82] MED *Relat.* culture, kidney, serum; epithelium, eye, liver, lens. *Vbs.* find. *Exp.* lung tissue (cf. electron microscopy, electron microscope), human lung (cf. hela cell, lymph node).

lymphocyte :: [115 contexts, frequency rank 84] MED *Relat.* transformation, percentage, nucleus, index, population. *Vbs.* label. *Fam.* lymph, lymph node, lymphatic, lymphoid.

marrow :: [189 contexts, frequency rank 46] MED *Relat.* culture; metamyelocyte, experiment, age, suspension, spleen, kidney, serum, liver. *Vbs.* irradiate, shield, obtain, appear. *Exp.* bone marrow (cf. peripheral blood, growth hormone)

material :: [125 contexts, frequency rank 78] MED *Relat.* control, examination; procedure; . *Vbs.* contain.

mean :: [92 contexts, frequency rank 104] MED *Relat.* method; model, half life, measure. *Vbs.* study. *Exp.* mean concentration (cf. acid metabolism, acid composition)

measurement :: [93 contexts, frequency rank 103] MED *Relat.* deficiency, property, analysis, reduction; rise; evaluation, estimation, determination, record, detection.

mechanism :: [136 contexts, frequency rank 72] MED *Relat.* role, process; factor; investigation. *Vbs.* discuss, investigate, explain.

membrane :: [88 contexts, frequency rank 107] MED *Relat.* dirofilaria, endothelium, lamella, granule, infiltration. *Exp.* cell membrane (cf. basement membrane, flow rate), basement membrane (cf. connective tissue, type ii).

metabolism :: [105 contexts, frequency rank 91] MED *Relat.* synthesis, concentration, content; size, depletion, phospholipid, utilization, composition, glucose, mobilization. *Exp.* protein metabolism (cf. zona glomerulosa, folic acid), acid metabolism (cf. mean concentration, folic acid), carbohydrate metabolism (cf. protein metabolism, blood glucose). *Fam.* metabolic.

method :: [298 contexts, frequency rank 23] MED *Relat.* test; mean, procedure, technique. *Vbs.* use, describe, make, modify, improve, show, present, outline, consider, apply.

mice :: [76 contexts, frequency rank 118] MED *Relat.* mouse; woman, infant, animal, rat, dog; puppy, hypophysectomized, sheep, rabbit. *Vbs.* treat.

mouse :: [75 contexts, frequency rank 119] MED *Relat.* mice; dog, rat; c3h, sarcoma, hypophysectomized, hybrid, rabbit, male.

nephrectomy :: [75 contexts, frequency rank 119] MED *Relat.* calvaria, vagotomy, hospitalization, yr, x-irradiation, parathyroidectomy, myocardium, transection, pregnancy, ligation. *Vbs.* follow. *Fam.* nephrectomized.

number :: [183 contexts, frequency rank 49] MED *Relat.* injection; concentration, activity; amount. *Vbs.* increase, reduce, observe, compose.

observation :: [181 contexts, frequency rank 50] MED *Relat.* pattern, difference; case, result, study; evaluation, analysis, data, finding. *Vbs.* indicate, suggest, show, make.

operation :: [85 contexts, frequency rank 109] MED *Relat.* part, surgery; therapy, treatment, procedure; repair, amputation, resection, malformation. *Vbs.* follow, perform. *Fam.* operative.

output :: [87 contexts, frequency rank 108] MED *Relat.* volume, tension; citrate, sensitivity, po2, spray, s.c., calcium, glucose, production. *Vbs.* increase.

oxygen :: [89 contexts, frequency rank 106] MED *Relat.* temperature; flow, blood; sodium, retention, admixture, air, po2. *Vbs.* breathe. *Exp.* oxygen tension (cf. carbon dioxide, blood pressure), oxygen consumption (cf. carbon dioxide, coronary flow).

part :: [90 contexts, frequency rank 105] MED *Relat.* antibody, operation; area, role; phenomenon, psychology, resistance. *Vbs.* play.

patient :: [883 contexts, frequency rank 2] MED *Relat.* woman, disease, day, effect, treatment, result, group, child, study, case. *Vbs.* treat, receive, occur, excrete, study, die, find, use, show, select, suffer, perform. *Exp.* cancer patient (cf. survival time, joint deformity), adult patient (cf. protein excretion, liver cell).

pattern :: [171 contexts, frequency rank 56] MED *Relat.* observation; type, study, change, difference; sign, relationship, feature, characteristic. *Vbs.* show, indicate.

period :: [227 contexts, frequency rank 34] MED *Relat.* rate; degree, duration, stage, time. *Vbs.* occur, follow, study, find, remain, die. *Exp.* year period (cf. survival rate, hormone therapy)

phase :: [81 contexts, frequency rank 113] MED *Relat.* period, stage; .

plasma :: [94 contexts, frequency rank 102] MED *Relat.* lens, liver; value, marrow, blood, serum; . *Exp.* plasma calcium (cf. plasma phosphate, vitamin d), plasma concentration (cf. urine volume, sodium intake), plasma phosphate (cf. plasma calcium, parathyroid hormone), plasma cell (cf. blood stream, surface activity).

preparation :: [90 contexts, frequency rank 105] MED *Relat.* extract; fraction; trace, spray, vasopressin. *Vbs.* purify.

presence :: [153 contexts, frequency rank 65] MED *Relat.* relationship; absence. *Vbs.* show, suggest, demonstrate, influence.

pressure :: [286 contexts, frequency rank 24] MED *Relat.* level, rate; serum, tension, obstruction, volume, flow. *Vbs.* increase, associate, raise. *Exp.* blood pressure (cf. oxygen tension, carbon dioxide), fluid pressure (cf. carbon dioxide, blood flow), pressure curve (cf. right ventricle, left ventricle).

problem :: [127 contexts, frequency rank 76] MED *Relat.* data, child, form; symptom, feature, disorder, disturbance. *Vbs.* present, learn.

procedure :: [154 contexts, frequency rank 64] MED *Relat.* therapy, treatment, technique, method; examination, criterion, operation, surgery. *Vbs.* describe, carry.

process :: [136 contexts, frequency rank 72] MED *Relat.* mechanism; condition, phenomenon, structure. *Vbs.* involve.

property :: [85 contexts, frequency rank 109] MED *Relat.* measurement, extract; fraction, component; nature, determinant, enzyme, composition, distribution, constituent.

protein :: [212 contexts, frequency rank 36] MED *Relat.* dna; activity, acid, growth, hormone; molecule, analysis, antigen. *Vbs.* contain. *Exp.* protein metabolism (cf. zona glomerulosa, folic acid), lens protein (cf. acid metabolism, lens epithelium), serum protein (cf. inclusion body, dilution curve), protein fraction (cf. insoluble protein, ionic strength), protein component (cf. wuchereria bancrofti, skin reaction), protein excretion (cf. adult patient, filtration rate), insoluble protein (cf. protein fraction, m urea).

rat :: [331 contexts, frequency rank 22] MED *Relat.* group; rabbit, day, kidney, infant, mice, dog, mouse, animal. *Vbs.* treat, give, expose, determine, study, receive, produce, feed, maintain, fast. *Exp.* rat kidney (cf. folic acid, testosterone propionate)

rate :: [387 contexts, frequency rank 15] MED *Relat.* result, response, increase; effect, change, level; pressure, time, value, concentration. *Vbs.* increase, decrease, find, reduce, induce, follow, determine. *Exp.* growth rate (cf. growth retardation, folic acid), flow rate (cf. coronary flow, body temperature), survival rate (cf. radiation therapy, survival time), heart rate (cf. fluid pressure, pressure curve), filtration rate (cf. vitamin d, urine volume).

ratio :: [115 contexts, frequency rank 84] MED *Relat.* rise; weight, excretion, content, level, concentration; capacity, clearance, glucose. *Vbs.* increase, make, find, decrease.

reaction :: [245 contexts, frequency rank 32] MED *Relat.* growth, test; child, increase, effect, response; antibody, antigen, relationship. *Vbs.* produce, show, study, involve, inhibit. *Exp.* stress reaction (cf. blood viscosity, compound lipid), skin reaction (cf. protein component, cobalt chloride).

reduction :: [97 contexts, frequency rank 99] MED *Relat.* measurement, correlation, alteration; relationship, increase, decrease, rise; variation, fall. *Vbs.* mark. *Fam.* reduce.

relationship :: [139 contexts, frequency rank 71] MED *Relat.* presence; reaction, response, function, pattern; correlation, reduction, contact, relation. *Vbs.* indicate, exist, establish.

report :: [94 contexts, frequency rank 102] MED *Relat.* experiment, author; data; trial, paper, communication, review. *Vbs.* publish. *Exp.* case report (cf. intra-arterial infusion, age group)

response :: [389 contexts, frequency rank 14] MED *Relat.* result, rate, increase; study, change, effect; function, value, treatment, reaction. *Vbs.* show, suggest, relate, obtain, make, evoke.

result :: [446 contexts, frequency rank 9] MED *Relat.* case, patient, study, effect; group, data, finding, observation, rate, response. *Vbs.* suggest, indicate, show, obtain, give, discuss, compare, report, yield, interpret, demonstrate, make.

rise :: [119 contexts, frequency rank 82] MED *Relat.* ratio, decrease; concentration, difference, increase; glucose, retention, fall, elevation, reduction. *Vbs.* show, mean, give, follow.

role :: [128 contexts, frequency rank 75] MED *Relat.* mechanism; factor; pathogenesis, influence, relation, importance, part. *Vbs.* play, appear.

serum :: [150 contexts, frequency rank 67] MED *Relat.* value, pressure, hormone, blood, level, marrow; liver, plasma, lens, lense. *Vbs.* increase, use, find, demonstrate. *Exp.* serum protein (cf. inclusion body, dilution curve)

stage :: [126 contexts, frequency rank 77] MED *Relat.* area; development, period; condition, course, phase. *Exp.* stage iv (cf. initial value, total estrogen)

state :: [89 contexts, frequency rank 106] MED *Relat.* bleeding, pig. *Vbs.* lead.

strain :: [178 contexts, frequency rank 52] MED *Relat.* species, phage, virus. *Vbs.* isolate, infect, use, compare. *Exp.* mycoplasma strain (cf. cell culture, tissue culture), cell strain (cf. control group, total content).

structure :: [118 contexts, frequency rank 83] MED *Relat.* process, area, body; system; layer, surface. *Vbs.* observe.

study :: [626 contexts, frequency rank 4] MED *Relat.* case, effect; patient; treatment, rate, response, result, observation, change. *Vbs.* show, make, indicate, use, suggest, detail, describe, confirm, carry, perform, evaluate, determine.

subject :: [139 contexts, frequency rank 71] MED *Relat.* rat, group, patient, child; breast, dog, man, woman. *Vbs.* receive, cool.

surgery :: [85 contexts, frequency rank 109] MED *Relat.* operation; treatment, therapy, procedure; element, diagnostic, scan. *Vbs.* follow. *Fam.* surgical.

symptom :: [98 contexts, frequency rank 98] MED *Relat.* syndrome, problem, finding, lesion; concept, sign, manifestation.

syndrome :: [206 contexts, frequency rank 38] MED *Relat.* lesion, group, disease, case, type; symptom. *Vbs.* describe.

synthesis :: [162 contexts, frequency rank 58] MED *Relat.* formation; content, concentration; component, transport, production, molecule, metabolism. *Vbs.* undergo.

system :: [160 contexts, frequency rank 60] MED *Relat.* area, structure, organ. *Exp.* nervous system (cf. heart disease, inclusion disease)

technique :: [188 contexts, frequency rank 47] MED *Relat.* therapy, test, method; dog, data, analysis, procedure. *Vbs.* use, study, describe, demonstrate.

temperature :: [97 contexts, frequency rank 99] MED *Relat.* oxygen; perfusion. *Exp.* body temperature (cf. extracorporeal circulation, flow rate)

tension :: [124 contexts, frequency rank 79] MED *Relat.* fluid; pressure, blood; availability, washing, vein, ph, output, po2. *Vbs.* measure, increase. *Exp.* oxygen tension (cf. carbon dioxide, blood pressure), surface tension (cf. surface activity, total lipid).

test :: [284 contexts, frequency rank 25] MED *Relat.* therapy, method; response; observation, analysis, reaction, technique. *Vbs.* use, indicate, give, utilize, follow, carry.

therapy :: [256 contexts, frequency rank 28] MED *Relat.* test; response, treatment; procedure, operation, drug, chemotherapy, dose, administration. *Vbs.* use, respond, follow, remain, receive, combine. *Exp.* radiation therapy (cf. survival rate, cancer chemotherapy), steroid therapy (cf. inclusion disease, cancer chemotherapy), hormone therapy (cf. intra-arterial infusion, steroid therapy), corticosteroid therapy (cf. connective tissue, plasma concentration). *Fam.* therapeutic.

time :: [204 contexts, frequency rank 40] MED *Relat.* injection, period, day; level, rate, group; month, age. *Vbs.* reach, prolong, increase, appear. *Exp.* survival time (cf. cancer chemotherapy, cancer patient), time constant (cf. action potential, flow rate).

tissue :: [350 contexts, frequency rank 18] MED *Relat.* disease; cell; resistance, serum, lens, liver, lesion, tumor, cancer, growth. *Vbs.* explant, obtain, decrease, recover, observe. *Exp.* tissue culture (cf. human lung, electron microscopy), lung tissue (cf. electron microscopy, electron microscope), human tissue (cf. inclusion body, human cell), connective tissue (cf. basement membrane, corticosteroid therapy).

treatment :: [341 contexts, frequency rank 19] MED *Relat.* response; effect, result, study, case, patient; surgery, diagnosis, administration, therapy. *Vbs.* follow, consist, stop, result, require, make, give, continue, combine. *Exp.* cortisone treatment (cf. total estrogen, sodium intake)

tumor :: [260 contexts, frequency rank 26] MED *Relat.* lesion, growth, cancer; effect, tissue, disease; sarcoma, carcinoma. *Vbs.* grow, use, suggest, produce. *Exp.* tumor growth (cf. body growth, tenuazonic acid), human tumor (cf. ,).

type :: [249 contexts, frequency rank 30] MED *Relat.* disease, defect, group, change, case; syndrome, pattern, feature, line, form. *Vbs.* show, observe, find, classify. *Exp.* type ii (cf. inclusion body, basement membrane), cell type (cf. vit d, survival rate).

value :: [202 contexts, frequency rank 42] MED *Relat.* response, increase, effect, rate, level, concentration; serum, rise, decrease. *Vbs.* increase, reach, find, mean. *Exp.* initial value (cf. stage iv, side effect)

ventricle :: [79 contexts, frequency rank 115] MED *Relat.* curve, artery; exercise, atrium, cistern, sinus, dimension, myocardium. *Exp.* left ventricle (cf. right ventricle, valve replacement), right ventricle (cf. left ventricle, dilution curve). *Fam.* ventricular.

virus :: [87 contexts, frequency rank 108] MED *Relat.* dna, growth, antigen, strain; particle. *Vbs.* induce. *Fam.* viral.

volume :: [180 contexts, frequency rank 51] MED *Relat.* excretion, flow; concentration, pressure; detection, incidence, output. *Vbs.* decrease, reduce, increase, implant. *Exp.* stroke volume (cf. valve replacement, blood volume), blood volume (cf. stroke volume, blood glucose), urine volume (cf. sodium intake, plasma concentration).

weight :: [172 contexts, frequency rank 55] MED *Relat.* dose; concentration, content; survival, rise, ratio, size, incidence. *Vbs.* increase, find. *Exp.* body weight (cf. kidney weight, dna content), kidney weight (cf. dna content, body weight), birth weight (cf. blood glucose, dna content).

woman :: [119 contexts, frequency rank 82] MED *Relat.* subject; patient, child; hour, year, mice, rabbit, male, female, mother. *Vbs.* find.

year :: [103 contexts, frequency rank 93] MED *Relat.* woman; child, patient, day; week, month, hour. *Vbs.* age, occur, follow. *Exp.* year period (cf. survival rate, hormone therapy)

6

CORPORA TREATED

6.1 ADI

Name	:	ADI
Size	:	39 kilobyte
Documents	:	82 (Average = 67 words)
Words	:	5470
Unique words	:	1473
Source	:	IR testbed (ftp'ed from ftp.cs.cornell)
Description	:	Library science abstracts
Queries	:	35 (Average = 16 words)

Sample Text:

- a new automatic method is presented for the comparison of two-dimensional line patterns . retrieval applications include the matching of chemical structures, the comparison of syntactically analyzed excerpts extracted from documents and search requests, and the matching of document identifications consisting of twodimensional graphs with query identifications .
- /letters/ journals, which developed out of the /letters to the editor/ section of research journals, provide rapid dissemination of results which are judged likely to have marked effects on the work of a substantial number of people . this is accomplished by keeping the communications brief, reviewing them promptly, and making use of rapid publication methods .

- it is proposed to develop a primary publication procedure which in addition to publishing the journal, records data needed for secondary publishing, and storage and retrieval purposes . the limitations of typography and the requirement for a recording procedure which identifies content or item function are stated . the problems of complex symbol representation are posed .
- an operational definition is attempted for the new composite discipline /information science/. the approach is based on the physics and psychology of messages. to include all steps in the /information transfer chain/ the definition must remain general . a probable limit of specificity is suggested in the /final/ definition offered .

Sample Queries :

- What problems and concerns are there in making up descriptive titles? What difficulties are involved in automatically retrieving articles from approximate titles? What is the usual relevance of the content of articles to their titles?
- How can actually pertinent data, as opposed to references or entire articles themselves, be retrieved automatically in response to information requests?
- What is information science? Give definitions where possible.
- Educational and training requirements for personnel in the information field. Possibilities for this training. Needs for programs providing this training.
- Describe information retrieval and indexing in other languages. What bearing does it have on the science in general?
- The use of abstract mathematics in information retrieval, e.g. group theory.

ADI (39K) : SEXTANT results, 50 most frequent words

word [*Contexts*]	*Groups of closest words. (See page 50)*
system [128]	technique | method program service thesaurus science
information [65]	data | plan program approach | basis access
index [37]	basis awareness | access copy record method retrieval
science [33]	health library service | feature book | center technique
center [31]	approach | network | procedure thesaurus extension service
program [30]	training | personnel prospect thesaurus procedure | search
retrieval [29]	copy apparatus | transmission format | dissemination
method [28]	technique | area selector index mode console dissemination
service [27]	science extension feature technique | center program
document [24]	record term | procedure
data [20]	character information | center dissemination approach
library [20]	catalogue | science | center | dissemination
format [19]	medicu | transmission type | form retrieval access
procedure [19]	approach | availability distribution | program center
technique [19]	vocabulary | method network service correlation book
thesaurus [18]	health | plan program center | retrieval system
problem [17]	process organization | technique | system
computer [16]	tape | deck | index dissemination
copy [16]	retrieval | index service | system
evaluation [16]	implementation | activity procedure | application function
access [15]	reproduction | format | index retrieval | information
analysis [15]	activity | organization device technique | evaluation
work [15]	scientist personnel | center | procedure organization
dissemination [14]	center retrieval data method term
organization [14]	device | analysis | problem procedure method work
role [14]	content | program system
need [13]	challenge education | system
plan [13]	thesaurus | information
area [12]	method
effort [12]	system
journal [12]	report
approach [11]	design center | procedure | extension network facility
communication [11]	information procedure
definition [11]	challenge education personnel | program
exchange [11]	amount storage | procedure technique center service
form [11]	format
reader [11]	logic | information system
report [11]	aerospace facility | journal | program | information
tape [11]	readout | application | output computer | method
function [10]	transformation | generator | evaluation procedure
study [10]	program | information
vocabulary [10]	technique term | system
basis [9]	awareness | index | information
catalogue [9]	feature | volume | record library
console [9]	display | utility | method
device [9]	organization | analysis | procedure work program
logic [9]	reader | application | evaluation procedure center
personnel [9]	scientist technician | education | challenge | prospect
research [9]	data
result [9]	system

ADI. Query Experiments Results

See Page 105

	ADI						
	base	DOC	SEXT	stem	fam	S+fam	S+f+stem
	PRECISION						
Recall: 10	0.667	0.670	0.655	0.699	0.667	0.655	0.687
Recall: 20	0.598	0.635	0.614	0.630	0.598	0.614	0.636
Recall: 30	0.517	0.544	0.532	0.527	0.524	0.532	0.539
Recall: 40	0.473	0.475	0.480	0.484	0.473	0.485	0.492
Recall: 50	0.458	0.470	0.464	0.468	0.458	0.464	0.478
Recall: 60	0.314	0.354	0.330	0.369	0.314	0.330	0.377
Recall: 70	0.223	0.292	0.237	0.288	0.223	0.237	0.290
Recall: 80	0.216	0.286	0.219	0.269	0.216	0.219	0.277
Recall: 90	0.184	0.244	0.186	0.236	0.184	0.186	0.244
Average	0.405	0.441	0.413	0.441	0.406	0.413	0.447
Better	---	18	7	14	1	7	16
Same	---	2	22	9	34	22	7
Worse	---	15	6	12	0	6	12
	RECALL						
At 5 docs:	0.26	0.27	0.26	0.27	0.26	0.26	0.27
At 10 docs:	0.20	0.20	0.21	0.21	0.20	0.21	0.22
At 15 docs:	0.17	0.16	0.17	0.17	0.17	0.17	0.18
At 20 docs:	0.14	0.14	0.14	0.14	0.14	0.14	0.14
At 25 docs:	0.12	0.12	0.12	0.13	0.12	0.12	0.12
Better at 15	---	5	3	3	0	3	6
Same at 15	---	22	32	31	35	32	27
Worse at 15	---	8	0	1	0	0	2

ADI --- BEST IMPROVEMENTS (*see page 105*)

Base Query	*Augmented Query*	*change*
system incorporate multiprogramming remote station information retrieval extent future	system incorporate multiprogramming remote station information retrieval **copy** extent future	0.437 to 0.538
obtain large volume high speed customer usable information retrieval output	obtain large volume high speed customer usable information retrieval **copy** output	0.093 to 0.170
retrieval system provide automate transmission information user distance	retrieval copy system provide automate transmission information user distance	0.532 to 0.608
educational training requirement personnel information field possibility training need program provide training	educational **educate** training **program** requirement personnel information field possibility training **program** need program **training** provide training **program**	0.826 to 0.895

ADI --- WORST RESULTS

Base Query	*Augmented Query*	*change*
technique machine match machine search system code match	technique **vocabulary** machine match machine search system code match	0.417 to 0.402
government support agency project information dissemination	government support agency project **deal** information dissemination	0.385 to 0.330

ADI. First-Pass Thesaurus. *See Page 131.*

access :: [15 contexts, frequency rank 16] ADI *Relat.* information, index, retrieval, format; reproduction.

analysis :: [15 contexts, frequency rank 16] ADI *Relat.* organization, evaluation; index, technique, retrieval; device, activity.

center :: [31 contexts, frequency rank 5] ADI *Relat.* program; storage, work, service, thesaurus, facility, data, procedure, network, approach. *Vbs.* specialize.

computer :: [16 contexts, frequency rank 15] ADI *Relat.* index, system; deck, tape. *Vbs.* use.

data :: [20 contexts, frequency rank 11] ADI *Relat.* technique, procedure; retrieval, method, center, information; research, approach, dissemination, character.

dissemination :: [14 contexts, frequency rank 17] ADI *Relat.* center, data, method, retrieval.

document :: [24 contexts, frequency rank 10] ADI *Relat.* term, record. *Vbs.* educate.

evaluation :: [16 contexts, frequency rank 15] ADI *Relat.* procedure, analysis; index, center, retrieval; function, logic, application, activity, implementation.

format :: [19 contexts, frequency rank 12] ADI *Relat.* retrieval; access, form, type, transmission, medicus.

index :: [37 contexts, frequency rank 3] ADI *Relat.* computer, evaluation, analysis, method, record, citation, retrieval, copy, access, basis. *Vbs.* produce, make.

information :: [65 contexts, frequency rank 2] ADI *Relat.* report, reader, access, basis, plan, character, approach, data, program, service. *Vbs.* process. *Exp.* information science (cf. information center,), information center (cf. information science,).

library :: [20 contexts, frequency rank 11] ADI *Relat.* service, science; catalogue.

method :: [28 contexts, frequency rank 8] ADI *Relat.* retrieval; index; organization, data, dissemination, console, mode, selector, area, technique. *Vbs.* use, code.

organization :: [14 contexts, frequency rank 17] ADI *Relat.* work, analysis; method, problem; device.

plan :: [13 contexts, frequency rank 18] ADI *Relat.* information, thesaurus. *Vbs.* index.

problem :: [17 contexts, frequency rank 14] ADI *Relat.* organization, process. *Vbs.* discuss.

procedure :: [19 contexts, frequency rank 12] ADI *Relat.* technique; program, center; application, evaluation, distribution, test, user, network, approach.

program :: [30 contexts, frequency rank 6] ADI *Relat.* center; term, report, service, search, procedure, thesaurus, personnel, prospect, training. *Vbs.* accredit.

retrieval :: [29 contexts, frequency rank 7] ADI *Relat.* method; index; search, activity, aid, dissemination, format, transmission, copy.

science :: [33 contexts, frequency rank 4] ADI *Relat.* program, center; media, book, feature, technique, service, library, thesaurus, health.

service :: [27 contexts, frequency rank 9] ADI *Relat.* program, center; information, science; copy, reflection, facility, correlation, technique, feature. *Vbs.* index, orient, give.

6.2 AI

Name	:	AI
Size	:	2.8 Megabyte
Documents	:	3254 (Average = 119 words)
Words	:	387 K
Unique words	:	25 K
Source	:	CLARIT test corpus
Description	:	Abstracts from AI articles

Sample Text:

- This decade has seen the appearance of several attempts to apply artificial intelligence (AI) to problems in computational fluid dynamics (CFD). The author proposes an approach for analyzing such AI/CFD systems, applies this analysis to four first-generation systems, and uses the results to assess the progress that has been made and highlights the remaining challenges. These first AI/CFD systems demonstrate that present AI technology can be successfully applied to well-formulated problems that are solved by means of classification or selection of pre-enumerated solutions (as opposed to construction, where solutions must be synthesized) . . .
- Up to now most knowledge based systems for differential diagnosis haven't got enough performance to be accepted as a useful tool, since large amounts of data are required. In this study a system is described, which is programmed in Pascal and stores information in a network-like pointer-structure. The following four levels have been defined: the physical access, single data and their relations, sets of information and their projections as well as the user-interface. Information from medical literature has been used to create the knowledge base. Due to the high efficiency of the system even a personal-computer is sufficient to make large amounts of medical information accessible for differential diagnosis . . .
- The View Creation System (VCS) is an expert system that engages a user in a dialogue about the information requirements for some application, develops an entity-relationship model for the user's database view, and then converts the E-R model to a set of fourth normal form relations. This paper describes the knowledge base of VCS. That is, it presents a formal methodology, capable of VCS. That is, it presents a formal methodology, capable of mechanization as a computer program, for accepting requirements from a user, identifying and resolving inconsistencies, redundancies . . .

AI (2800K) : SEXTANT results, 50 most frequent words

word [*Contexts*]	*Groups of closest words. (See page 50)*
system [14892]	author | strategy task procedure set mechanism | paper form
model [2234]	problem | information algorithm environment | result
problem [2188]	approach model application method technique | tool concept
knowledge [2099]	approach | technique | language environment concept | set
approach [1916]	technique method | problem knowledge process tool
technique [1769]	method | approach | tool | knowledge technology problem
author [1738]	paper | result method user tool technique | application
application [1709]	approach | method implementation | analysis architecture
design [1558]	analysis | technology | solution representation type
process [1490]	method | approach | tool | strategy rule aspect
method [1451]	technique | approach | tool process algorithm problem
expert [1431]	model | technique | representation information result
intelligence [1356]	software ai | computer simulation control issue support
development [1295]	implementation | problem architecture representation
tool [1295]	technique | technology method | approach | methodology
program [1269]	algorithm | result technology rule architecture base
language [1216]	knowledge | approach tool methodology structure | software
control [1213]	tool | controller analysis representation simulation
analysis [1163]	design | representation application approach description
information [1153]	model | rule description tool | application analysis
base [1083]	data_base | concept representation program methodology tool
environment [1043]	knowledge | technology approach technique model problem
data [1023]	technique | tool program environment approach | concept
representation [989]	description | form approach | analysis rule concept
technology [985]	tool technique | approach environment design method
rule [953]	method | approach technique information strategy program
algorithm [927]	method | technique program | model architecture problem
structure [926]	architecture | method concept language approach framework
network [881]	tool process algorithm | analysis mechanism level method base
result [862]	technique approach program methodology | model procedure
paper [824]	author | article | prototype approach | example technique
interface [805]	environment | methodology simulation base task technique
architecture [765]	structure methodology implementation type tool algorithm
concept [702]	technique approach method methodology | tool structure
set [687]	number type representation description knowledge | concept
computer [677]	intelligence | development methodology knowledge machine
management [645]	diagnosis analysis type design task | procedure plan
function [637]	concept aspect tool | component structure feature
software [626]	intelligence | language concept | approach data process
data_base [625]	base | representation tool | simulation set component
strategy [624]	technique | method approach rule concept process procedure
user [620]	technology | tool program application development author
example [618]	aspect | framework study experience representation case
reason [616]	inference aspect | knowledge concept theory procedure
theory [607]	logic representation aspect | method analysis concept
task [589]	capability technique problem activity concept | procedure
operation [582]	task | plan data program application representation behavior
issue [578]	aspect | requirement concept methodology problem | role
feature [576]	concept aspect type methodology function approach technique
level [572]	aspect | type component | description domain area approach

AI. Semantic Clusters.

See Page 126

idea *as* discussion	overview
impact *as* enhancement	improvement
impact *as* implication	potential
impact *as* potential	implication
implementation *as* architecture	representation
implementation *as* description	representation
implementation *as* development	design
implementation *as* methodology	architecture tool concept
implementation *as* simulation	design
implication *as* direction	trend
implication *as* impact	potential trend
implication *as* potential	impact
implication *as* progress	trend
improvement *as* enhancement	increase
improvement *as* increase	progress
improvement *as* objective	goal
improvement *as* progress	trend
improvement *as* trend	advantage
industry *as* company	plant
information *as* data-base	representation
information *as* description	representation structure
information *as* process	technique approach
information *as* rule	model
input *as* sub-system	facility
interaction *as* dialogue	communication
interest *as* attention	emphasis
interest *as* emphasis	attention
interest *as* utilization	acceptance
interface *as* base	program
interface *as* environment	tool knowledge
interpretation *as* inference	reason
introduction *as* background	overview
introduction *as* consideration	discussion
introduction *as* discussion	overview consideration idea
introduction *as* review	overview discussion
introduction *as* variety	overview
issue *as* aspect	requirement
issue *as* development	problem
issue *as* requirement	aspect
kind *as* amount	variety
kind *as* class	type
kind *as* variety	class amount
knowledge *as* concept	approach technique method
knowledge *as* environment	language
knowledge *as* method	approach technique
knowledge *as* set	rule
knowledge *as* technique	approach
language *as* environment	knowledge tool
language *as* methodology	tool software architecture

AI. Semantic Clusters. Two-word Terms.

See Page 145

image-analysis *as* computer-vision	vision-system
image-analysis *as* feature-extraction	image-data
image-analysis *as* hybrid-approach	image-data
image-analysis *as* image-understanding	vision-system computer-vision
image-analysis *as* object-recognition	vision-system computer-vision
image-analysis *as* scene-analysis	image-understanding
image-data *as* feature-extraction	image-analysis
image-data *as* hybrid-approach	image-analysis
image-sequence *as* image-understanding	machine-vision vision-system
image-sequence *as* scene-analysis	image-understanding image-analysis
image-understanding *as* computer-vision	vision-system
image-understanding *as* image-analysis	vision-system computer-vision
image-understanding *as* machine-vision	vision-system computer-vision
image-understanding *as* scene-analysis	image-sequence image-analysis
inference-engine *as* data-structure	user-interface production-system
inference-engine *as* expert-knowledge	production-rule
inference-engine *as* knowledge-acquisition	knowledge-representation natural-language
inference-engine *as* management-system	user-interface
inference-engine *as* production-rule	expert-knowledge production-system
inference-engine *as* user-interface	knowledge-representation
inference-mechanism *as* domain-knowledge	production-rule
inference-mechanism *as* expert-control	dynamic-system
inference-mechanism *as* production-rule	expert-knowledge inference-engine
inference-mechanism *as* rule-base	domain-knowledge
inference-method *as* knowledge-structure	ai-system
inference-rule *as* order-logic	horn-clause knowledge-elicitation
information-system *as* computer-system	decision-support user-interface
information-system *as* decision-support	natural-language knowledge-representation
information-system *as* knowledge-acquisition	decision-support natural-language
information-system *as* knowledge-representation	knowledge-base
information-system *as* management-system	user-interface
information-system *as* natural-language	knowledge-representation knowledge-base
information-system *as* user-interface	decision-support natural-language
information-technology *as* project-management	human-factor ai-technique
intelligence-approach *as* ai-method	speech-act
intelligence-technique *as* decision-support	natural-language
intelligence-technique *as* power-plant	process-control
intelligence-technique *as* process-control	knowledge-acquisition decision-support
intelligence-technique *as* system-design	software-system
interface-design *as* dynamic-system	man-machine-system human-computer-inter
interface-design *as* future-development	software-design
interface-design *as* man-machine-system	dynamic-system
interface-design *as* software-design	future-development
internal-representation *as* tutorial-system	subject-matter
knowledge-acquisition *as* computer-system	decision-support process-control
knowledge-acquisition *as* control-system	knowledge-representation knowledge-base
knowledge-acquisition *as* decision-support	knowledge-representation knowledge-base
knowledge-acquisition *as* inference-engine	knowledge-representation user-interface

AI. First-Pass Thesaurus. *See Page 131.*

algorithm :: [927 contexts, frequency rank 26] AI *Relat.* program, technique, method, problem, model, approach; architecture, concept. *Vbs.* use, present, describe, develop, process, base, propose, give, learn, call, discuss, show. *Exp.* control algorithm (cf. expert control, computer technology), rete algorithm (cf. real time application, multilayer perceptron), search algorithm (cf. intelligence problem, heuristic search), heuristic algorithm (cf. vlsi design, hybrid approach).

analysis :: [1163 contexts, frequency rank 18] AI *Relat.* approach, application, process, design, model; study, structure, description, representation. *Vbs.* use, perform, base, present, provide, model, require, process, lead, engineer, detail, describe. *Exp.* data analysis (cf. consultation system, case study), image analysis (cf. vision system, computer vision), protocol analysis (cf. user requirement, design decision), system analysis (cf. human activity, requirement specification), comparative analysis (cf. set theory, design system), scene analysis (cf. image understanding, image analysis), risk analysis (cf. management support, process operation), decision analysis (cf. business application, future research), analysis system (cf. software tool, performance evaluation).

application :: [1709 contexts, frequency rank 8] AI *Relat.* technique, approach; model; result, analysis, implementation, technology, process, method, tool. *Vbs.* discuss, describe, develop, engineer, use, consider, present, involve, select, process, illustrate, give. *Exp.* ai application (cf. ai system, future development), application area (cf. performance evaluation, computer architecture), application domain (cf. network management, expert knowledge), intelligence application (cf. computer technology, ai method), potential application (cf. ai research, project management), computer application (cf. agv system, project management), business application (cf. decision analysis, data management), application system (cf. human computer interface, blackboard architecture), application program (cf. level language, office automation), space application (cf. project management, fault tree).

approach :: [1916 contexts, frequency rank 5] AI *Relat.* model, knowledge, problem, technique; concept, tool, algorithm, application, process, method. *Vbs.* use, present, propose, base, describe, take, provide, integrate, discuss, show, develop, adopt. *Exp.* intelligence approach (cf. ai method, performance evaluation), system approach (cf. requirement specification, factory automation), novel approach (cf. order logic, microcomputer system), hybrid approach (cf. image analysis, image data), connectionist approach (cf. intelligence system, present state), alternative approach (cf. future research, hybrid system), ai approach (cf. timetable preparation, abstraction level).

architecture :: [765 contexts, frequency rank 32] AI *Relat.* concept; representation, approach, algorithm, application, tool, model, structure; implementation, methodology. *Vbs.* describe, use, base, propose, present, discuss, provide, process, give, implement, distribute, develop. *Exp.* system architecture (cf. cad system, support system), blackboard architecture (cf. real time control, human computer interface), computer architecture (cf. performance evaluation, ai language), software architecture (cf. real time system, application system), network architecture (cf. pattern classification, decision problem).

base :: [1083 contexts, frequency rank 20] AI *Relat.* tool, program, representation; interface, data-base. *Vbs.* use, contain, consist, construct, build, represent, create, describe, store, establish, develop, update. *Exp.* knowledge base (cf. knowledge representation, knowledge acquisition), rule base (cf. domain knowledge, expert knowledge), data base (cf. cad system, system architecture), information base (cf. problem area, novice user).

6.3 AIDS

Name	:	AIDS
Size	:	2.8 megabyte
Documents	:	2344 (Average = 195 words)
Words	:	458K
Unique words	:	22K
Source	:	IR testbed developed by Dr. Hersh (Univ Oregon)
Description	:	AIDS abstracts
Queries	:	75 (Average = 8 words)

Sample Text:

- Post-jejunoileal-bypass hepatic disease. Its similarity to alcoholic hepatic disease. The authors studied serial hepatic biopsies of five patients who developed hepatic failure following jejunoileal bypass for extreme obesity, with autopsies of two. The hepatic histologic changes included centrilobular or focal alcoholic hyalin, intrasinusoidal collagenosis, fatty hydropic degeneration, and neutrophilic infiltrate. At least two of the patients were abstinent from alcohol, both prior to and after the surgical procedures. The others, after the bypass procedures, had reduced alcohol consumption from previous levels. All patients developed hepatic failure and histologically . . .
- Laboratory diagnosis and monitoring of rheumatologic diseases. Improved laboratory investigative techniques now foster an increased clinical interest in and awareness of the rheumatologic disease. This review is a discussion of the relevance of laboratory tests used in the more common rheumatologic disorders and of their role in both the diagnosis and assessment of these diseases from the standpoint of the practising clinician.
- Recent trends in breast-cancer incidence and mortality in relation to changes in possible risk factors. Breast cancer incidence and mortality in England and Wales and the United States increased between 1950 and 1973, mainly in women aged between 45 and 64 years. These increases appeared to be partly cohort-specific, beginning with cohorts born around 1899, and partly cross-sectional, beginning in the mid-1960s. In both countries, cohort-specific decreases in fertility paralleled the cohort-specific increases in breast cancer rates and may, at least in part, have been responsible for them. Changes in other factors, such as age at menarche and menopause, use of rauwolfia derivatives and oestrogens, consumption of fat and meat, and breast cancer treatment were considered in relation to the cross-sectional increases in breast cancer rates. On the evidence available, it was not certain that any of these could explain the breast cancer increases.

Sample Queries :

- patient with mycosis fungoides, wish to assess treatment options
- malnutrition secondary to malabsorption related to crohn's disease
- stroke treatment with calcium channel blockers
- I would like to find out the relationship between Reiter's Syndrome and vasculitis for diagnostic and therapeutic reasons.
- natural history of supraventricular tachycardia in pregnancy
- Bartter's Disease

AIDS (2800K) : SEXTANT results, 50 most frequent words

word [*Contexts*]	*Groups of closest words. (See page 50)*
patient [6950]	case | study effect therapy result | rate woman day child
effect [2204]	response | therapy change result increase rate | study
treatment [1920]	therapy | study dose chemotherapy | drug administration
disease [1904]	infection disorder | complication syndrome case response
study [1820]	trial result treatment | data therapy finding change effect
therapy [1694]	treatment | chemotherapy | dose effect group
group [1626]	therapy | case rate effect dose result response
rate [1554]	risk result | incidence time increase difference | effect
level [1553]	concentration | value | response activity increase rate
cell [1441]	response result | study serum infection tissue factor group
response [1362]	effect | activity level concentration change result
result [1257]	study rate data finding | effect response case treatment
infection [1082]	disease | complication | case meningitis antibody
syndrome [1062]	disease | symptom disorder complication | case infection
case [930]	patient result | incidence disease treatment data year child
cancer [905]	carcinoma | disease tumor | age woman lesion failure
trial [883]	study | treatment evaluation therapy | result program
change [866]	increase abnormality | reduction effect response difference
factor [841]	mechanism | result effect activity time | response agent
activity [815]	response | level effect function | concentration increase
pressure [814]	flow | index value concentration volume level | risk plasma
concentration [812]	level | value response | increase dose rate content
difference [745]	improvement reduction increase | rate change value comparison
function [723]	response abnormality dysfunction activity damage disease flow
dose [718]	therapy treatment administration regimen | dosage
increase [710]	reduction | change decrease difference improvement | rate
risk [706]	incidence | rate mortality | death survival frequency
test [697]	method study response finding | evaluation examination
day [669]	week | month | year minimum | hour time | period
blood [661]	plasma | concentration serum value | exposure mortality
infarction [649]	neuropathy event necrosis damage outcome stroke death ischemia
drug [629]	agent | treatment | chemotherapy result control medication
time [610]	rate day duration | week month dose year | course result
survival [601]	outcome | prognosis duration mortality | follow-up month
antibody [576]	antigen | infection concentration change virus result
flow [569]	pressure | volume perfusion function hypertension
control [566]	mortality | woman result management recovery response therapy
data [566]	result finding | study evidence report case investigation
value [556]	level | concentration | ratio difference measurement
symptom [523]	sign disorder complication finding | abnormality diagnosis
period [520]	day duration week | year survival month course hour | rate
system [510]	test mechanism activity method response | result technique
finding [501]	feature | result data abnormality | symptom change study
chemotherapy [491]	therapy | radiotherapy treatment | cmf administration
year [488]	month | week day | child case data period | infant time
exposure [484]	administration transfusion | treatment response value effect
failure [480]	toxicity reaction dysfunction injury damage necrosis function
neuropathy [477]	nerve | dysfunction | damage symptom complication
diagnosis [466]	evidence evaluation cause management symptom case | sign
analysis [464]	study result data comparison finding | measurement evaluation

AIDS. Semantic Clusters.

See Page 126

illness *as* manifestation	problem sign
illness *as* problem	sign
implication *as* hypothesis	support
implication *as* significance	benefit
implication *as* theory	hypothesis
importance *as* benefit	significance
importance *as* characteristic	need
importance *as* effectiveness	benefit
importance *as* need	characteristic
importance *as* relevance	significance strategy
importance *as* significance	benefit
improvement *as* benefit	correlation
improvement *as* correlation	difference decrease
improvement *as* decrease	reduction difference increase
improvement *as* difference	change
improvement *as* fall	reduction decrease
incidence *as* association	risk
incidence *as* frequency	risk mortality
incidence *as* mortality	risk survival
incidence *as* occurrence	frequency development
incidence *as* prevalence	frequency
incidence *as* risk	rate
incidence *as* survival	risk
increase *as* change	rate effect
increase *as* concentration	level
increase *as* decrease	reduction difference improvement
increase *as* difference	change rate
increase *as* improvement	reduction change difference
increase *as* level	rate effect
increase *as* rate	effect level
increase *as* reduction	change difference
increase *as* rise	decrease
index *as* pressure	concentration level
indomethacin *as* cyclosporine	nimodipine ml
indomethacin *as* nimodipine	ml
indomethacin *as* propranolol	cyclosporine
infant *as* adult	child individual
infant *as* child	woman year day
infant *as* individual	adult
infant *as* man	child woman year subject
infant *as* neonate	adult
infant *as* subject	child woman month
infarction *as* damage	neuropathy death
infarction *as* event	complication
infarction *as* stroke	event
infection *as* complication	disease
influence *as* relationship	role association difference
information *as* criterion	feature
information *as* experience	data report
information *as* report	data
infusion *as* administration	dose transfusion therapy
infusion *as* injection	administration

AIDS. Semantic Clusters. Two-word Terms.

See Page 145

immunodeficiency-syndrome *as* aids-related-complex	peripheral-neuropathy
immunodeficiency-syndrome *as* blood-product	blood-transfusion hepatitis-b
immunodeficiency-syndrome *as* factor-viii	blood-product hepatitis-b
immunodeficiency-syndrome *as* hepatitis-b	blood-transfusion liver-disease
immunodeficiency-syndrome *as* peripheral-neuropathy	nervous-system
immunosuppressive-agent *as* animal-model	plasma-exchange
immunosuppressive-agent *as* calcium-channel-blocker	channel-blocker
immunosuppressive-agent *as* cytomegalovirus-infection	plasma-exchange
immunosuppressive-agent *as* multiple-sclerosis	plasma-exchange
immunosuppressive-agent *as* oral-corticosteroid	plasma-exchange
immunosuppressive-therapy *as* cytomegalovirus-disease	cytomegalovirus-infection
immunosuppressive-therapy *as* initial-diagnosis	natural-history
indomethacin-treatment *as* bartter-syndrome	preterm-infant
indomethacin-treatment *as* distress-syndrome	ductus-arteriosus preterm-infant
indomethacin-treatment *as* indomethacin-group	indomethacin-therapy
indomethacin-treatment *as* indomethacin-therapy	ductus-arteriosus
indomethacin-treatment *as* preterm-infant	ductus-arteriosus
initial-therapy *as* adult-patient	antibiotic-therapy
initial-therapy *as* salvage-therapy	adjuvant-therapy
iron-deficiency *as* erythroid-progenitor	cell-cycle
iron-deficiency *as* laboratory-parameter	hemoglobin-concentration
kda-protein *as* protein-synthesis	amino-acid
laboratory-test *as* ct-scan	stroke-patient
laboratory-test *as* laboratory-data	time-interval
liquid-chromatography *as* escherichia-coli	amino-acid
liquid-chromatography *as* ril--beta	escherichia-coli amino-acid
listeria-monocytogene *as* brain-abscess	laboratory-finding blood-culture
listeria-monocytogene *as* laboratory-finding	brain-abscess adult-patient
listeria-monocytogene *as* literature-review	brain-abscess case-report
liver-biopsy *as* alcoholic-cirrhosis	shunt-surgery
liver-biopsy *as* factor-viii	immunodeficiency-syndrome
liver-biopsy *as* replacement-therapy	factor-viii
liver-cirrhosis *as* liver-function	liver-disease
liver-cirrhosis *as* plasma-level	blood-cell
liver-cirrhosis *as* potassium-excretion	sodium-excretion
liver-damage *as* renin-angiotensin-aldosterone-system	kallikrein-excretion
liver-disease *as* blood-transfusion	risk-factor
liver-disease *as* hepatitis-b	risk-factor blood-transfusion
liver-disease *as* liver-function	liver-cirrhosis serum-albumin
liver-disease *as* risk-factor	heart-disease
liver-enzyme *as* hellp-syndrome	platelet-count
liver-function *as* kallikrein-excretion	creatinine-clearance
liver-function *as* liver-cirrhosis	liver-disease
liver-function *as* serum-albumin	liver-disease
liver-function *as* weight-loss	creatinine-clearance
lobe-epilepsy *as* absence-spell	epileptic-patient seizure-disorder
lobe-epilepsy *as* panic-attack	panic-disorder
lobe-epilepsy *as* seizure-disorder	epileptic-patient
long-term-follow-up *as* radionuclide-ventriculography	time-interval
long-term-survival *as* channel-blocker	beta-blocker
long-term-survival *as* host-disease	marrow-transplantation

AIDS. First-Pass Thesaurus. *See Page 131.*

activity :: [815 contexts, frequency rank 20] AIDS-ABSTRACTS *Relat.* factor, pressure, concentration; rate, effect, level, response; function. *Vbs.* increase, reduce, show, measure, find, correlate, suppress, involve, restore, remain, relate, inhibit. *Exp.* nk activity (cf. t cell, prostaglandin e2), disease activity (cf. fluid volume, plasma aldosterone), enzyme activity (cf. odcase mrna, calf serum), physical activity (cf. serum cholesterol, low back pain).

antibody :: [576 contexts, frequency rank 35] AIDS-ABSTRACTS *Relat.* infection; virus, antigen. *Vbs.* develop, use, produce, demonstrate, react, neutralize, detect, suggest, occur, block, test, observe. *Exp.* antibody response (cf. cell wall, serum antibody), antibody level (cf. thromboplastin time, lupus anticoagulant), serum antibody (cf. antibody level, antibody response), antibody titer (cf. time interval, disease control), lymphocyte antibody (cf. disease process, airflow obstruction), igg antibody (cf. cell wall, cytomegalovirus antibody), cytomegalovirus antibody (cf. igg antibody, cell wall). *Fam.* antigen.

blood :: [661 contexts, frequency rank 30] AIDS-ABSTRACTS *Relat.* artery, fluid, exposure, value, serum, plasma. *Vbs.* transfuse, receive, increase, freeze, use, match, isolate, give, condition. *Exp.* blood pressure (cf. heart rate, blood flow), blood flow (cf. blood pressure, heart rate), blood transfusion (cf. blood product, risk factor), blood product (cf. blood transfusion, cytomegalovirus infection), blood cell (cf. blood transfusion, blood product), blood loss (cf. confidence interval, blood transfusion), peripheral blood (cf. bone marrow, cell line), blood sample (cf. average time, lactate concentration), blood gas (cf. hemodynamic change, hemoglobin concentration), blood culture (cf. h influenzae, brain abscess).

cancer :: [905 contexts, frequency rank 16] AIDS-ABSTRACTS *Relat.* disease; failure, ascites, lesion, tumor, carcinoma. *Vbs.* advance, develop, treat, smoke, detect, use, review, randomize, induce, increase. *Exp.* breast cancer (cf. adjuvant chemotherapy, response rate), lung cancer (cf. risk factor, heart disease), cancer patient (cf. antibiotic therapy, combination therapy), cancer treatment (cf. median follow up, multicenter trial).

case :: [930 contexts, frequency rank 15] AIDS-ABSTRACTS *Relat.* group, disease, treatment, result, patient; woman, child, year, incidence, data. *Vbs.* report, present, occur, treat, describe, find, observe, review, show, follow, study, proven. *Exp.* case report (cf. literature review, transverse myelitis), case history (cf. stroke patient, pregnancy induced hypertension).

cell :: [1441 contexts, frequency rank 10] AIDS-ABSTRACTS *Relat.* tissue. *Vbs.* infect, culture, use, contain, circulate, obtain, multinucleate, transfer, mediate, grow, express, show. *Exp.* blood cell (cf. blood transfusion, blood product), t cell (cf. graft survival, mycosis fungoides), mast cell (cf. plasma cell, tumor cell), cell line (cf. peripheral blood, ay27 cell), b cell (cf. t cell, igg synthesis), cell carcinoma (cf. case report, cell line), cell membrane (cf. intracellular ca2, liver enzyme), tumor cell (cf. mast cell, lymph node), plasma cell (cf. median survival, radiation therapy), cell population (cf. dna synthesis, suspension culture).

6.4 ANIMALS

Name	:	ANIMALS
Size	:	1.2 Megabyte
Documents	:	756 (Average = 260 words)
Words	:	200K
Unique words	:	18000
Source	:	Grolier's Encyclopedia animal articles
Description	:	Long articles were truncated
Queries	:	none

From the encyclopedia were drawn 756 animal articles:
aardwolf, abalone, Abyssinian cat, accentor, addax, adder, Afghan hound, agama, agouti, Airedale terrier, albatross, alderfly, alligator, alpaca, ammonite, amoeba, amphioxus, amphipod, anchovy, . . . , wildcat, wildebeest, wolf, wolverine, wombat, wood swallow, woodcock, woodpecker, worm, wrasse, wren, wryneck, yak, Yorkshire terrier, zebra.

Sample Text:

- The Abyssinian is a medium-sized cat with a triangular face, large pointed ears, lean body, and long tail. The almond-shaped eyes are hazel to orange, and the fine, short-haired coat ranges from light brown to silver. Each hair is ticked, or banded, with darker brown, gray, or black. The tail and ears are darker toward the tip. The red Abyssinian is a recognized breed.
- The agouti, genus Dasyprocta, is a rodent belonging to the family Dasyproctidae, order Rodentia. About 24 species exist. Agoutis are about 61 cm (24 in) long, have short tails (10-35, or 2/5 to 1 2/5 in), and weigh up to 3 kg (7lb). The coarse, glossy coat ranges from pale orange to shades of brown and near-black; the underparts are whitish. The body is slender with a high, muscular rump well adapted for running. The front paws have five claws and the hind have three much thicker claws. Agoutis live from southern Mexico to southern Brazil. They flourish in cool, damp lowland forests, grasslands, and brush, and feed on leaves, fruit, nuts, and roots. They dig burrows in which they raise litters of two to four.
- The addax, Addax nasomaculatus, is the single species of addax among many antelope species in the family Bovidae of the order Artiodactyla. It is found only in the Sahara, where it was once widespread. The plump, short-legged addax is more than 1.8 m (6 ft) long and about 1 m (39 in) tall at the shoulders, and it weighs up to 120 kg (265 lb). Both sexes have long horns that are ringed and screw-shaped. The coat is gray to white, with a black skullcap and facial markings. The broad hooves are an adaptation for travel on desert sands. Addaxes usually get water only from the plants they eat, but in captivity they drink large amounts. Unable to flee hunters as speedily as some other antelopes do, the addax is an endangered species.
- The Appaloosa is a breed of horse developed by the nez perce Indians. In October 1877, Chief Joseph and his Nez Perce band surrendered to the U.S. Army and were exiled to Oklahoma, taking with them 1,100 of their carefully bred Appaloosa horses.

ANIMALS (1200K) : SEXTANT results, 50 most frequent words

word [*Contexts*]	*Groups of closest words. (See page 50)*
species [1404]	bird fish | family group form | animal insect range snake
fish [771]	animal species bird | form snake | insect | group water
bird [622]	species | fish animal | snake insect form | mammal duck
water [533]	sea area region | coast | forest ocean part fish form lake
egg [517]	nest female | male larva insect day sperm form | adult
animal [510]	fish bird | form insect snake | species group larva mammal
body [445]	head tail bill | foot leg scale animal form feather ear
family [401]	species | genera subfamily range form | order variety size
tail [385]	bill wing | leg head | body coat ear foot | toe fur
range [350]	variety | size family species north color form | forest
nest [345]	burrow egg | hole cell animal range species insect bird live
form [337]	group fish animal insect species type | bird reptile snake
dog [333]	terrier | horse animal year bird bear | wolf type form nest
insect [315]	form animal | bird fish species reptile snake egg plant
group [288]	form species | reptile order animal snake variety kind
area [285]	forest water region | ocean sea part | country coast plain
food [283]	prey animal organism | material crustacean feed egg bird part
fin [277]	scale teeth appendage leg spine | head ear organ eye bone
name [269]	characteristic | study number stage term bear scale feed form
year [269]	month | litter form number dog individual length day egg male
snake [266]	lizard | fish animal bird viper group form reptile mammal
head [265]	bill tail body | wing back coat color fin ear eye
system [264]	structure tract | group tube form gland pair behavior lack
length [263]	weight | size | height maturity shoulder | speed
member [263]	bird group species animal form fish reptile monkey | mammal
leg [251]	wing tail bill neck | foot toe claw limb fin | antenna
male [251]	female | egg ft lb individual bird species | animal year
centimeter [250]	kg mm | ft hand neck | lb length maturity | tail leg
part [249]	area region | coast end surface island side front | water
coat [245]	fur | plumage hair | color feather patch | stripe marking
number [234]	size population | variety horse group year name species
region [234]	sea | forest | area water | africa coast europe plain
kg [232]	mm | centimeter | turkey | length individual young speed
female [230]	male | egg | larva individual deer bird sex bee layer day
eye [223]	ear scale opening | fin head wing organ toe teeth leg
variety [214]	range type group animal kind | form number variation feed
forest [197]	africa | region lake | area coast south plain north-america
scale [193]	skull fin eye feather | skin teeth plate shell body skeleton
color [186]	plumage | green marking coat blue shade | patch coloration
organ [186]	structure viper fin eye opening | teeth cell function ear
bill [184]	tail | leg neck head claw | foot body wing face beak
teeth [183]	fin spine bill | scale jaw pair eye wing claw foot
pair [173]	end | segment opening teeth base plate | toe structure wing
ear [172]	eye | muzzle tail fin head toe | bill wing nose leg
side [171]	front top | back stripe band | bone surface marking | end
structure [169]	organ system | end feature type body scale pad pair skull
larva [168]	animal egg beetle female worm form type stage | colony brood
size [167]	length | number | weight range temperature | shape
feed [165]	variety form kind bear predator food name land egg | move
mammal [157]	rodent animal lizard | snake form crab life reptile bird deer

ANIMALS. Semantic Clusters.

See Page 126

insect *as* animal	bird fish species
insect *as* bird	fish species
insect *as* fish	species
insect *as* form	animal bird fish species
insect *as* snake	form animal bird fish
kg *as* centimeter	length
larva *as* female	egg
larva *as* type	form
leg *as* bill	tail
leg *as* claw	bill foot toe
leg *as* foot	wing tail bill
leg *as* limb	wing foot
leg *as* neck	bill foot
leg *as* toe	tail foot
leg *as* wing	tail bill foot
length *as* height	weight maturity shoulder speed
length *as* kg	centimeter
length *as* maturity	weight height speed centimeter kg proportion
length *as* shoulder	weight
length *as* speed	weight height maturity
length *as* temperature	size
length *as* weight	size
lizard *as* mammal	snake animal
lizard *as* rodent	mammal frog
lizard *as* snake	animal insect
male *as* bird	species
male *as* female	egg bird
mammal *as* animal	bird
mammal *as* lizard	snake
mammal *as* rodent	lizard
mammal *as* snake	animal form bird
material *as* grass	seed
material *as* pollen	grass wood
monkey *as* vulture	warbler
mouth *as* pouch	opening
mouth *as* spine	claw fin
nest *as* hole	burrow
nest *as* range	species
number *as* group	species
number *as* size	length
number *as* variety	group
organ *as* eye	fin
organ *as* opening	eye
pair *as* end	structure
part *as* area	water
part *as* end	side
part *as* front	surface side
part *as* region	area water
part *as* surface	side
plumage *as* band	fur stripe patch marking
plumage *as* color	coat
plumage *as* fur	coat

ANIMALS. Semantic Clusters. Two-word Terms.

See Page 145

hind-leg *as* order-orthoptera	front-leg
hind-limb *as* canine-teeth	hind-leg
left-ventricle *as* body-surface	book-lung
left-ventricle *as* mantle-tissue	right-ventricle internal-organ
left-ventricle *as* right-ventricle	blood-vessel
life-cycle *as* adult-stage	adult-form
life-cycle *as* animal-kingdom	horseshoe-crab
life-cycle *as* digestive-system	adult-form digestive-tract
marine-water *as* atlantic-coast	shallow-water
marine-water *as* coral-reef	shallow-water marine-fish
marine-water *as* coral-snake	pit-viper
marine-water *as* marine-fish	atlantic-coast coral-reef
marine-water *as* order-perciformes	atlantic-coast
marine-water *as* sea-level	shallow-water pit-viper
marine-water *as* total-length	sea-level
marine-water *as* warmer-region	coral-snake
nervous-system *as* blood-vessel	body-cavity
nervous-system *as* book-lung	compound-eye
nervous-system *as* compound-eye	book-lung
nervous-system *as* head-region	sense-organ mantle-tissue
nervous-system *as* mantle-tissue	head-region sense-organ
nervous-system *as* pit-organ	digestive-tract
nervous-system *as* sense-organ	head-region mantle-tissue compound-eye
nervous-system *as* vertebrate-animal	blood-vessel digestive-tract
norway-rat *as* barbary-ape	kangaroo-rat body-length tree-shrew life-span
norway-rat *as* filter-feeder	food-chain
norway-rat *as* kangaroo-rat	barbary-ape canine-teeth
order-artiodactyla *as* cat-family	order-carnivora weasel-family
order-artiodactyla *as* order-insectivora	order-rodentia order-carnivora
order-artiodactyla *as* order-perissodactyla	order-rodentia order-carnivora
order-artiodactyla *as* order-rodentia	order-perissodactyla order-insectivora
order-carnivora *as* cat-family	weasel-family order-artiodactyla
order-carnivora *as* family-anatidae	family-bovidae
order-carnivora *as* order-artiodactyla	family-bovidae
order-carnivora *as* order-insectivora	order-artiodactyla order-rodentia order-primate
order-carnivora *as* order-perissodactyla	order-artiodactyla order-rodentia
order-carnivora *as* sea-lion	family-bovidae
order-carnivora *as* weasel-family	cat-family family-bovidae
order-passeriformes *as* order-rodentia	thrush-family coral-snake
order-passeriformes *as* sea-level	pit-viper marine-water
pit-organ *as* gill-opening	ear-opening sense-organ
pit-organ *as* prey-animal	body-temperature
pit-viper *as* coral-snake	marine-water
pit-viper *as* salt-water	game-fish
pit-viper *as* sea-level	order-passeriformes rain-forest marine-water
plant-material *as* aquatic-animal	vegetable-matter rock-crevice
plant-material *as* cup-shaped-nest	vegetable-matter rock-crevice
plant-material *as* rock-crevice	vegetable-matter cup-shaped-nest
plant-material *as* vegetable-matter	rock-crevice cup-shaped-nest aquatic-animal

ANIMALS. First-Pass Thesaurus. *See Page 131.*

animal :: [510 contexts, frequency rank 6] ANIMALS *Relat.* bird, species, fish; prey, mammal, larva, group, snake, insect, form. *Vbs.* feed, use, eat, domesticate, live, prey, kill, classify, call, know, graze, enable. *Exp.* prey animal (cf. pit organ, body temperature), aquatic animal (cf. surface area, aquatic life), animal kingdom (cf. phylum arthropoda, phylum mollusca), vertebrate animal (cf. blood vessel, digestive tract), land animal (cf. african elephant, life span).

area :: [285 contexts, frequency rank 16] ANIMALS *Relat.* water; country, coast, sea, ocean, part, region, forest. *Vbs.* find, inhabit, live.

bird :: [622 contexts, frequency rank 3] ANIMALS *Relat.* fish, species; female, male, duck, mammal, form, insect, snake, animal. *Vbs.* know, find, live, relate, feed, raise, call, wad, use, perch, medium-size, fly. *Exp.* game bird (cf. family anatidae, game fish), water bird (cf. family anatidae, food item).

body :: [445 contexts, frequency rank 7] ANIMALS *Relat.* scale, foot, leg, bill, tail, head. *Vbs.* compress, cover, flatten, elongate, streamline, divide, round, hold, extend, consist. *Exp.* body temperature (cf. deg f, deg c), body cavity (cf. body wall, internal organ), body length (cf. shoulder height, barbary ape), body wall (cf. internal organ, body cavity), body weight (cf. body length, average weight), body surface (cf. book lung, surface area).

centimeter :: [250 contexts, frequency rank 25] ANIMALS *Relat.* length, kg; little, lb, neck, hand, ft, maturity, mm. *Vbs.* grow, reach, weigh, stand, measure, average, range, vary, mature, use, streak, extend.

coat :: [245 contexts, frequency rank 27] ANIMALS *Relat.* tail; color, feather, hair, plumage, fur. *Vbs.* vary, mark, cover.

color :: [186 contexts, frequency rank 35] ANIMALS *Relat.* coat; shade, coloration, blue, patch, marking, fur, plumage. *Vbs.* vary, change.

dog :: [333 contexts, frequency rank 13] ANIMALS *Relat.* year, horse, terrier. *Vbs.* breed, use, develop, weigh, know, hunt, work. *Exp.* prairie dog (cf. ground squirrel, vegetable matter), sled dog (cf. standard schnauzer, quarter horse).

egg :: [518 contexts, frequency rank 5] ANIMALS *Relat.* adult, sperm, day, form, insect, larva, male, female, nest. *Vbs.* lay, lie, incubate, fertilize, hatch, develop, produce, contain, spot, shed, release, deposit.

eye :: [223 contexts, frequency rank 31] ANIMALS *Relat.* leg, head; fin; toe, organ, opening, teeth, scale, ear. *Vbs.* set, locate, cover, situate, protect, degenerate, bear. *Exp.* compound eye (cf. nervous system, ear opening), eye socket (cf. front wing, mantle cavity).

family :: [401 contexts, frequency rank 8] ANIMALS *Relat.* species; order, nest, range, subfamily, genera. *Vbs.* belong, classify, comprise, own, contain, relate, place, occur, mingle, make, divide, consider. *Exp.* family bovidae (cf. order artiodactyla, africa south), weasel family (cf. order carnivora, cat family), cat family (cf. order carnivora, weasel family), thrush family (cf. order passeriformes, coral snake), family anatidae (cf. water bird, game bird), jack family (cf. spanish mackerel, ground squirrel), horse family (cf. order perissodactyla, africa south), family boidae (cf. hind limb, coral snake).

female :: [230 contexts, frequency rank 30] ANIMALS *Relat.* male; bird, egg; layer, bee, day, sex, deer, individual, larva. *Vbs.* incubate, lie, lay, hatch, give, build, bear, weigh, attract, use, retain, produce.

fin :: [277 contexts, frequency rank 18] ANIMALS *Relat.* head, leg; bone, organ, ear, spine, appendage, eye, teeth, scale. *Vbs.* pair, call, modify, join, consist, reduce, extend.

6.5 BASEBALL

Name	:	BASEBALL
Size	:	946 kilobyte
Documents	:	not divided into documents
Words	:	190,035
Unique words	:	16080
Source	:	downloaded from NetNews rec.baseball
Description	:	groups of submissions from 1991 and 1992
Queries	:	none

Sample Text:

- My question is -- what do I do with it? I would like to have a combination statistic that factors in defensive as well as offensive performance. Right now I am using Usefulness percentage: (Total Bases + Stolen Hits)/Plate Appearances. If someone could suggest a better formula I would greatly appreciate it.
- Gotta be careful here. In 1963, the NL *average* OBP was .307 and SLG was .364. Rose was above average in both categories. This year, it's more like .330/.390 in the AL, and Knoblauch doesn't look so hot.
- Announcing the r.s.bb net.politically.correct postseason awards vote! We'll show those BBWAA dweebs what it's all about! The deadline for voting is (as of now) October 12th (Sat) at midnight. Any votes after that *may* be accepted, but no guarantees.

  To clarify a couple of things - Pitchers definitely *are* eligible for ROY, MVP, LVP, etc. and I encourage you to vote for them if you think they're deserving of that award. I've included categories like Rookie pitcher of the . . .
- Morris had his shot. You claim that an exhausted and spent Pendleton should have defended this title the day after winning the NL West. Every other Braves regular sat out the game.
- Suppose that you could with 100 accuracy measure the category "stolen hit". That is an out that would not have been made by an average fielder with average effort. (I know, it's undefineable, but this is for a computer game).

  My question is -- what do I do with it? I would like to have a combination statistic that factors in defensive as well as offensive performance. Right now I am using Usefulness percentage: (Total Bases + Stolen Hits)/Plate Appearances. If someone could suggest a better formula I would greatly appreciate it.

  Usefulness has some problems (for instance, Great shortstops tend to steal more hits than great outfielders, at least in my game) but it is useful for comparing players at the same position.

  I would also like help in this area -- what do I do with catchers? Should I factor in their stolen hits as being thrown out * 3 / attempted steals?

  Thanks in advance.

BASEBALL (950K) : SEXTANT results, 50 most frequent words

word [*Contexts*]	*Groups of closest words. (See page 50)*
fan [517]	player play | manager series time number win pitcher people
player [429]	pitcher | play hitter fan number point manager time run
brave [357]	pirate | inning | people twin series run number win point
ball [281]	run pitcher | time season runner lot brave way number guy
pitcher [280]	hitter | player season time win | lot number run series
series [270]	play | brave world-series pitcher twin | baseball way jay
number [254]	time point brave pitcher run stat player thing | play jay
run [251]	hit ball lot brave point hr pitcher number | play player
play [242]	series | hit player win hitter chance call jay time way
hit [214]	run play average | hr inning hitter pitcher time shot guy
time [209]	number pitcher inning jay baseball way guy play ball
pitch [208]	start hitter pitcher brave series player time inning job
jay [205]	atlanta twin pirate | lot toronto time play blue-jay pitcher
season [185]	pitcher | way ball starter nlc player league run time brave
thing [183]	number brave atlanta manager part point pitcher player people
lot [180]	run pitcher jay ball | bit hitter time brave chance runner
people [178]	brave jay | pitcher player pirate thing guy time series
hitter [176]	pitcher | chance fielder bullpen play manager player
call [175]	play number | cbs zone pitch hit time difference week
guy [154]	umpire time ump brave twin | pitcher atlanta runner people
flag [150]	league people way anthem baseball jay time chance part brave
point [144]	number run pitcher player brave stat rest | win time way
base [142]	alomar situation seat performance inning cover play home run
baseball [140]	sport | time mistake hitter series chance defense atlanta
way [139]	runner manager chance time win twin season play series pitcher
manager [134]	hitter way brave player | record play number win thing
league [131]	season record deal anthem flag brook situation franchise
win [117]	pitcher play brave edge chance victory way pirate offense
inning [111]	brave | time morris rally hit starter | pitch season base
pirate [105]	brave | jay | toronto atlanta win expo pittsburgh | twin
night [103]	atlanta afternoon effect way play deal run inning twin hit
average [102]	hit total | situation rest avg gene level difference hitter
stat [99]	number point performance record | average void guy change
runner [96]	error | way situation ball bat | back hitter guy pitcher
twin [95]	jay blue-jay brave | series way nlc atlanta guy toronto
day [94]	start baseball candiotti record | time part couple fun
problem [94]	reason support deal league difference rest bonilla ab call
zone [92]	ground | profit range call umpire pitch fielder ball help
man [91]	book side atlanta situation thing padre manager owner son
reardon [88]	stanton jay bullpen brave pitcher work | inning ward
world-series [82]	ws pennant deal series nlc fun alc division | red-sox time
position [81]	lineup ability root career pitcher fault advantage statement
rest [81]	bat point hockey average win stuff way | pitcher manager
chance [80]	offense bullpen credit hitter overlap | atlanta break play
right [80]	part roster pitcher hitter chance pitch point situation win
value [80]	specialist lineup effect | impact point throw average
part [79]	cub right thing winner | point world-series pitcher homer
field [78]	fielder place corner | variability possibility range town
mistake [77]	prospect blunder baseball error | reason fact difference
name [77]	leader number time mistake chance | answer kirby vote thing

BASEBALL. Semantic Clusters.

See Page 126

inning *as* hit	pitch
jay *as* atlanta	pirate twin toronto
jay *as* blue-jays	twin
jay *as* lot	pitcher
jay *as* pirate	twin
jay *as* team	game
jay *as* time	team game pitcher
jay *as* toronto	pirate atlanta twin
jay *as* twin	pirate
lot *as* jay	team game time
lot *as* pitcher	year team game
lot *as* run	year ball
lot *as* runner	ball
manager *as* hitter	player play
number *as* player	team
number *as* point	run player
number *as* run	brave
number *as* stat	point
number *as* time	pitcher team
people *as* brave	team
people *as* guy	time
people *as* pirate	brave jay
pirate *as* atlanta	jay toronto twin
pirate *as* brave	team
pirate *as* jay	team
pirate *as* people	brave
pirate *as* pittsburgh	atlanta
pirate *as* toronto	jay atlanta twin
pirate *as* twin	brave jay
pitch *as* hit	ball
pitch *as* hitter	pitcher player
pitcher *as* brave	year team game
pitcher *as* hitter	team player
pitcher *as* player	year team game
pitcher *as* season	year
pitcher *as* team	game
pitcher *as* time	year team game
pitcher *as* year	team game
play *as* chance	hitter
play *as* hitter	player team
play *as* player	year game team
play *as* team	game
play *as* year	game team
player *as* fan	game
player *as* hitter	team pitcher play
player *as* number	team
player *as* pitcher	team year game brave
player *as* play	team year game
player *as* point	year number

BASEBALL. Semantic Clusters. Two-word Terms.

See Page 145

gold-glove *as* base-hit	center-field rickey-henderson
gold-glove *as* pennant-race	post-season regular-season
gold-glove *as* post-season	regular-season jeff-reardon
gold-glove *as* regular-season	post-season
gold-glove *as* rickey-henderson	league-average
home-plate *as* home-run	strike-zone
home-plate *as* left-field	bobby-cox
home-plate *as* regular-season	home-run post-season bobby-cox
home-run *as* baseball-game	jeff-reardon post-season home-team
home-run *as* home-plate	strike-zone
home-run *as* home-team	regular-season post-season
home-run *as* jeff-reardon	strike-zone post-season
home-run *as* post-season	jeff-reardon regular-season
home-run *as* regular-season	post-season home-team home-plate
home-run *as* strike-zone	world-series
home-run *as* world-series-game	strike-zone jeff-reardon regular-season
home-team *as* ball-game	bobby-cox
home-team *as* baseball-fan	baseball-game
home-team *as* baseball-game	post-season home-run jeff-reardon
home-team *as* bobby-cox	jeff-reardon
home-team *as* man-rotation	post-season bobby-cox
home-team *as* post-season	regular-season home-run bobby-cox jeff-reardon
home-team *as* regular-season	post-season home-run bobby-cox
jeff-reardon *as* baseball-game	home-run post-season
jeff-reardon *as* bobby-cox	strike-zone
jeff-reardon *as* home-run	strike-zone
jeff-reardon *as* pinch-hit	post-season
jeff-reardon *as* pinch-hitter	bobby-cox world-series-game
jeff-reardon *as* post-season	bobby-cox home-run
jeff-reardon *as* strike-call	strike-zone
jeff-reardon *as* terry-pendleton	strike-call
jeff-reardon *as* world-series-game	home-run pinch-hitter strike-zone
league-average *as* plate-appearance	pennant-race
league-average *as* rickey-henderson	gold-glove
pennant-race *as* baseball-game	post-season jeff-reardon
pennant-race *as* gold-glove	regular-season post-season
pennant-race *as* post-season	regular-season jeff-reardon
pennant-race *as* regular-season	post-season gold-glove
pennant-race *as* terry-pendleton	jeff-reardon
post-season *as* baseball-game	pennant-race home-run jeff-reardon home-team
post-season *as* bobby-cox	jeff-reardon
post-season *as* gold-glove	regular-season
post-season *as* home-team	regular-season bobby-cox home-run
post-season *as* jeff-reardon	bobby-cox home-run
post-season *as* man-rotation	bobby-cox home-team
post-season *as* pennant-race	regular-season gold-glove
post-season *as* pinch-hit	jeff-reardon
post-season *as* regular-season	bobby-cox home-run gold-glove home-team
regular-season *as* gold-glove	post-season

BASEBALL. First-Pass Thesaurus. *See Page 131.*

average :: [102 contexts, frequency rank 34] BASEBALL *Relat.* hit; oba, level, gene, rest, avg, situation, total. *Vbs.* bat, hit.

ball :: [282 contexts, frequency rank 7] BASEBALL *Relat.* pitcher; brave, year; season, lot, runner, run, hit, pitch. *Vbs.* hit, play, touch, throw, call, take, give, catch, break, bat. *Exp.* fly ball (cf. plate appearance, center field), ball game (cf. home team, pat border), ball hit (cf. fly ball, center field).

base :: [143 contexts, frequency rank 26] BASEBALL *Relat.* ball, lot; cover, inning, performance, seat, situation, alomar, runner. *Vbs.* load, steal, run, leave.

baseball :: [140 contexts, frequency rank 27] BASEBALL *Relat.* series, time; day, defense, mistake, sport. *Vbs.* play, watch, know, learn, follow. *Exp.* baseball game (cf. blue jay fan, home team), baseball fan (cf. atlanta fan, baseball game), baseball history (cf. pennant race, clutch hitter), baseball team (cf. white sox fan, flag upside).

brave :: [357 contexts, frequency rank 6] BASEBALL *Relat.* year, game, team; run, series, pitcher, inning, twin, people, pirate. *Vbs.* win, think, lose, root, put, play, pitch, make, look, know, come, take.

call :: [175 contexts, frequency rank 22] BASEBALL *Relat.* play; zone. *Vbs.* make, blow, miss, give, think.

day :: [94 contexts, frequency rank 38] BASEBALL *Relat.* time, baseball; candiotti, start.

fan :: [517 contexts, frequency rank 4] BASEBALL *Relat.* year; team, game; player. *Vbs.* 'm, think, win, celebrate, turn, touch, steal, reach, raise, put, make, look. *Exp.* jay fan (cf. toronto fan, alex nava), brave fan (cf. atlanta fan, pinch hitter), twin fan (cf. death wear, place team), pirate fan (cf. baseball fan, left field), yankee fan (cf. place team, red sox fan), baseball fan (cf. atlanta fan, baseball game), toronto fan (cf. baseball fan, jay fan), dodger fan (cf. pennant race, post season), atlanta fan (cf. tire chain, baseball fan).

flag :: [150 contexts, frequency rank 24] BASEBALL *Relat.* incident, country, league. *Vbs.* look, carry, sell, invert, hold, display. *Exp.* flag incident (cf. color guard, flag upside), flag upside (cf. flag incident, white sox fan).

game :: [771 contexts, frequency rank 1] BASEBALL *Relat.* play, fan, player, jay, pitcher, time, brave, team, year. *Vbs.* win, play, lose, watch, pitch, say, start, make, come, sit, enjoy, attend. *Exp.* baseball game (cf. blue jay fan, home team), game series (cf. man rotation, regular season), ball game (cf. home team, pat border).

guy :: [154 contexts, frequency rank 23] BASEBALL *Relat.* people; brave, pitcher, year, time; job, runner, twin, umpire. *Vbs.* say, give, decide.

hit :: [210 contexts, frequency rank 13] BASEBALL *Relat.* play, pitch, run; ball; shot, inning, average. *Vbs.* give, come. *Exp.* base hit (cf. center field, home run), pinch hitter (cf. jeff reardon, bobby cox), pinch hit (cf. division title, jeff reardon), ball hit (cf. fly ball, center field).

6.6 BROWN

Name	:	BROWN
Size	:	6.2 megabyte
Documents	:	1695
Sentences	:	51279
Words	:	1,000,000
Unique	:	46K
Source	:	Kucera
Description	:	Varied English text
Queries	:	none

Sample Text:

- the Fulton County Grand Jury said Friday an investigation of Atlanta 's recent primary election produced no evidence that any irregularities took place . the jury further said in term-end presentments that the City Executive Committee , which had over-all charge of the election , deserves the praise and thanks of the City of Atlanta for the manner in which the election was conducted .
- said Benedick : Lady Beatrice , have you wept all this while ? replied Beatrice : yea , and I will weep a while longer . the heavens refused to give up their weeping . the gallant company completed Act 4 and got through part of Act 5 . but the final scenes could not be played .
- bed slats were washed in alum water , legs of beds were placed in cups of kerosene , and all woodwork was treated liberally with corrosive sublimate , applied with a feather . kerosene was very effective in ridding pioneer homes of the pests . at times pioneer children got lice in their hair .
- we were forbidden to swing on the gates , lest they sag on their hinges in a poor-white-trash way , but we could stand on them , when they were latched , rest our chins on the top , and stare and stare , committing to memory , quite unintentionally , all the details that lay before our eyes . the street that is full now of traffic and parked cars then and for many years drowsed on an August afternoon in the shade of the curbside trees , and silence was a weight , almost palpable , in the air .

BROWN (6200K) : SEXTANT results, 50 most frequent words

word [*Contexts*]	*Groups of closest words. (See page 50)*
man [2383]	people time | year | day way woman thing work child boy
time [1425]	year man | day way people thing number | part work week
year [1394]	time day man | people program student way thing work number
number [968]	time result value part work year line people life form
day [955]	year time | man hour week night month people thing work
people [924]	man | thing child time woman year way life person | girl
work [899]	program study problem number year people time | man life
program [891]	interest | activity work year system | need state plan
system [890]	program problem | method interest work cost process value
area [742]	problem part point time value way | place life development
problem [713]	situation question value need change work | area aspect
member [701]	student people man group | president officer year life
state [660]	program people value interest year action part | service
school [657]	education people child service college | problem house
group [655]	life organization activity people problem program number year
place [655]	way thing part life | time world house day right country
house [609]	room home place people | man life school building time day
word [560]	thing time people | part picture change name year man god
child [553]	people student boy | school mother teacher man person woman
service [543]	interest | cost study organization life school education
interest [540]	activity value program | experience service life need
head [513]	arm way body foot back room side hand | thing man
night [490]	morning day | week afternoon hour way time meeting month year
end [488]	side part | year way place time country experience home life
face [473]	eye | figure body hand woman day people man way thing
law [471]	action right problem organization authority policy program fact
cost [467]	rate service value price activity industry tax amount problem
result [453]	number change value part | time difference year information
home [451]	house year morning day child service office place end number
study [446]	work course development information program service | change
action [443]	law state interest part decision effect activity stage problem
kind [437]	form sort part thing sense program value time type problem
need [436]	value experience problem program demand difference activity
water [433]	floor light body foot people place time service money glass
name [432]	time word place day number thing form year way figure
boy [430]	girl | child student people mother man woman | thing day
right [426]	law place way week mother road matter service idea back
development [419]	activity organization | study interest experience problem
mean [415]	thing form value problem result process way difference force
plan [415]	program problem | organization interest study report activity
student [415]	child year people woman boy member person girl | teacher
effort [401]	interest program activity work problem way | attention world
policy [401]	program government service law position problem industry
order [391]	level world position problem result organization law form
fact [386]	problem principle thing experience study factor difference
method [386]	technique condition system | way procedure result problem
idea [383]	problem right form interest program number sense value way
president [376]	student year people government member | leader mother child
company [365]	government service program business reason development
course [352]	study program service time | stage work interest class result

BROWN. Semantic Clusters.

See Page 126

importance *as* confidence	judgment
importance *as* difference	value
income *as* calendar	capital
income *as* expenditure	sale assistance
increase *as* difference	change
increase *as* dollar	measure
increase *as* level	rate
industry *as* market	sale
industry *as* sale	market product
information *as* data	evidence study
information *as* evidence	data
information *as* objective	purpose
information *as* statement	report
institution *as* agency	organization
institution *as* organization	leader activity
interest *as* activity	program need development
interest *as* development	service experience
interest *as* experience	value life need development
interest *as* need	program value life
interest *as* organization	activity service development
interest *as* value	form
investigation *as* application	analysis
investigation *as* examination	preparation
investigation *as* preparation	application examination
issue *as* decision	question
issue *as* question	problem
issue *as* statement	information
judgment *as* confidence	importance
knowledge *as* understanding	training
leader *as* institution	organization
leader *as* organization	activity group
leadership *as* agency	institution
leadership *as* party	leader
leadership *as* security	freedom
leg *as* arm	hand
length *as* distance	temperature
length *as* stage	range
length *as* temperature	distance
length *as* weight	size
level *as* condition	rate
level *as* increase	rate
level *as* order	position
level *as* position	value
life *as* experience	interest world value need
life *as* need	interest value
life *as* people	way year
life *as* thing	people way year

BROWN. Semantic Clusters. Two-word Terms.

See Page 145

identification-card *as* belgian-government	tax-payment
identity-crisis *as* behavior-pattern	identity-diffusion unwed-mother value-system
identity-crisis *as* childhood-experience	identity-diffusion unwed-mother family-life
identity-crisis *as* family-life	identity-diffusion unwed-mother
identity-crisis *as* identity-diffusion	unwed-mother family-life behavior-pattern
identity-crisis *as* time-diffusion	identity-diffusion unwed-mother family-life
identity-crisis *as* value-system	identity-diffusion unwed-mother behavior-pattern
identity-diffusion *as* behavior-pattern	unwed-mother identity-crisis value-system
identity-diffusion *as* childhood-experience	unwed-mother identity-crisis time-diffusion
identity-diffusion *as* family-life	unwed-mother
identity-diffusion *as* identity-crisis	unwed-mother time-diffusion childhood-experience
identity-diffusion *as* time-diffusion	unwed-mother identity-crisis family-life
identity-diffusion *as* value-system	unwed-mother identity-crisis behavior-pattern
image-intensifier *as* anode-voltage	fiber-plate fiber-coupler chromatic-aberration
image-intensifier *as* chromatic-aberration	anode-voltage heat-transfer internal-reflection
image-intensifier *as* fiber-coupler	fiber-plate anode-voltage lens-system
image-intensifier *as* internal-reflection	fiber-plate
image-intensifier *as* lens-system	fiber-plate fiber-coupler
image-regulus *as* curve-j	multiple-secant singular-line invariant-line
image-regulus *as* f-curve	multiple-secant singular-line f-point
image-regulus *as* f-fold-secant	f-line multiple-secant singular-line f-point
image-regulus *as* f-line	multiple-secant singular-line f-point f-curve
image-regulus *as* f-point	f-line multiple-secant singular-line f-curve
image-regulus *as* invariant-line	multiple-secant singular-line f-curve
image-regulus *as* line-involution	singular-line invariant-line curve-j f-curve
image-regulus *as* point-j	f-point curve-j
image-regulus *as* singular-line	multiple-secant invariant-line curve-j f-curve
impact-rate *as* flux-value	mass-threshold space-probe water-interest
impact-rate *as* mass-threshold	flux-value space-probe
impact-rate *as* retention-period	pilot-plant oxidation-pond
impact-rate *as* space-probe	mass-threshold flux-value
impact-rate *as* water-interest	flux-value
import-quotas *as* fiscal-uniformity	gentile-jewish-relation
import-quotas *as* gentile-jewish-relation	civilizational-crisis
incest-story *as* mine-warfare	over-all-concept
inch-square *as* keel-line	plaster-board paper-pattern
inch-square *as* pit-run-gravel	roof-beam
inch-square *as* plaster-board	press-clay paper-pattern
inch-square *as* press-clay	plaster-board paper-pattern
income-tax *as* calendar-year	fiscal-year
income-tax *as* du-pont	fiscal-year
income-tax *as* general-motors-stock	du-pont
income-tax *as* j-reorganization	market-value
income-tax *as* market-value	general-motors-stock du-pont
income-tax *as* tax-payment	calendar-year
increase-rate *as* beef-cattle	feed-efficiency foot-rot feed-conversion
increase-rate *as* body-weight	feed-efficiency
increase-rate *as* feed-conversion	feed-efficiency foot-rot body-weight beef-cattle
increase-rate *as* feed-efficiency	body-weight

BROWN. First-Pass Thesaurus. *See Page 131.*

area :: [747 contexts, frequency rank 12] BROWN *Relat.* part, problem; way, time; point. *Vbs.* cover, use, serve, select, provide, depress, spread, park, lie, hunt, show, separate. *Exp.* service area (cf. daylight hour, skywave signal), land area (cf. oxidation pond, world production), city area (cf. social class, missionary outreach), area coverage (cf. decomposition theorem, indian trade).

child :: [553 contexts, frequency rank 28] BROWN *Relat.* school; time, man, people; family, teacher, mother, woman, boy, student. *Vbs.* make, own, help, watch, want, teach, say, feel, expect, bring, use, tell.

day :: [952 contexts, frequency rank 6] BROWN *Relat.* man, time, year; month, week, night, hour. *Vbs.* come, spend, say, take, pass, affix, find, use, tell, remember, own, appear.

end :: [490 contexts, frequency rank 36] BROWN *Relat.* side; way, time, part; . *Vbs.* come, put, reach, give, begin.

eye :: [646 contexts, frequency rank 22] BROWN *Relat.* hand; arm, woman, girl, face. *Vbs.* close, look, own, open, turn, star, say, meet, fix, raise, lie, catch.

force :: [487 contexts, frequency rank 37] BROWN *Relat.* value, power; form; energy, strength, mean, change, pressure. *Vbs.* require, destroy, arm, use, join, take, hold, enter. *Exp.* force requirement (cf. business outlook, area coverage), task force (cf. minority group, state law).

form :: [617 contexts, frequency rank 23] BROWN *Relat.* time, number; view, kind, mean, information, interest, value. *Vbs.* take, file, match, appear, use, store, mark, give, exist, represent, relate, reduce. *Exp.* text form list (cf. information cell, dictionary form), text form (cf. information cell, dictionary form), dictionary form (cf. text form, information cell), form f (cf. equation f, chlorine atom), form j (cf. classroom teacher, j statement), human form (cf. natural world, human life), form list (cf. cell f, address f), existent form (cf. family life, rehabilitation program), art form (cf. stereo j, european nation).

group :: [656 contexts, frequency rank 20] BROWN *Relat.* member, life; program, people; world, organization, activity. *Vbs.* make, come, give, hold, take, meet, join, bring, belong. *Exp.* group j (cf. anti b activity, agglutinin activity), space group (cf. unit cell, oxygen atom), religious group (cf. religious belief, host country), minority group (cf. task force, wagon train), social group (cf. social control, out of town school), peer group (cf. social behavior, childhood experience), muscle group (cf. pull exercise, visual representation), ivy group (cf. oxygen atom, training institution), interest group (cf. social class background, social responsibility), group member (cf. group control, mate selection).

hand :: [723 contexts, frequency rank 14] BROWN *Relat.* people; girl, side, body, eye, arm. *Vbs.* take, hold, put, set, own, throw, come, wave, wash, shake, extend, bring. *Exp.* left hand (cf. right hand, front porch), right hand (cf. left hand, pool owner), firm hand (cf. city employes, gold phone).

head :: [510 contexts, frequency rank 33] BROWN *Relat.* room, side, body; hand; back, arm. *Vbs.* shake, come, turn, lift, hold, lower, put, mount, make, grow, carry, bury. *Exp.* head start (cf. apprentice program, flower garden), head back (cf. pull exercise, dressing gown), design head (cf. piece j, press clay), screw head (cf. bar spacer, proper position), piston head (cf. combustion chamber, cylinder volume).

6.7 CACM

Name	:	CACM
Size	:	1.3 megabyte
Documents	:	3204 (Average = 60 words)
Words	:	193K
Unique words	:	9.7K
Source	:	IR testbed (ftp'ed from ftp.cs.cornell)
Description	:	Computer Science abstracts
Queries	:	64 (Average = 20 words)

Sample Text:

- Preliminary Report-International Algebraic Language
- Extraction of Roots by Repeated Subtractions for Digital Computers
- Secant Modification of Newton's Method
- LEM 1, Small Size General Purpose Digital Computer Using Magnetic (Ferrite) Elements

  The paper examines some of the questions of development and construction of a general purpose digital computer using contactless magnetic (ferrite) and capacitive "DEZU" (long duration capacitive memory) elements, developed at the Laboratory of Electrical Modeling VINITYI AN SSSR, under the supervision of Professor L.I. Gutenmacher.
- Glossary of Computer Engineering and Programming Terminology
- The Problem of Programming Communication with Changing Machines A Proposed Solution Part 2
- Proposal for an UNCOL
- Two Square Root Approximations

Sample Queries :

- What articles exist which deal with TSS (Time Sharing System), an operating system for IBM computers?
- I am interested in articles written either by Prieve or Udo Pooch
- I'm interested in mechanisms for communicating between disjoint processes, possibly, but not exclusively, in a distributed environment. I would rather see descriptions of complete mechanisms, with or without implementations, as opposed to theoretical work on the abstract problem. Remote procedure calls and message-passing are examples of my interests.
- SETL, Very High Level Languages

CACM (1300K) : SEXTANT results, 50 most frequent words

word [*Contexts*]	*Groups of closest words. (See page 50)*
system [2421]	algorithm | technique method | procedure | result
program [1583]	method | time procedure process | result function
method [1533]	technique algorithm | program problem system | model result
algorithm [1487]	method | technique | procedure program problem | system
problem [1180]	algorithm procedure method | technique | application
language [1124]	technique | application model algorithm procedure problem
time [951]	number program | technique procedure | method result
technique [832]	method procedure | algorithm | model scheme language system
number [797]	set | time | size structure function technique problem data
computer [775]	data | process number technique approach | model
structure [713]	representation | model technique operation number type
function [704]	program number structure data | set procedure time algorithm
data [636]	computer | information technique result time | procedure
paper [594]	method algorithm model result | application procedure program
result [556]	technique data program solution analysis method procedure
procedure [542]	technique | method algorithm | scheme problem program
set [538]	number | result procedure value solution application
model [531]	concept analysis approach | structure method language
analysis [517]	model result | structure application approach procedure
solution [489]	result procedure method technique | set approximation
process [407]	procedure technique program | data model computer structure
equation [394]	function | expression solution | model formula operation
table [387]	property structure set data | manipulation size number type
application [384]	example implementation language | design technique problem
code [340]	technique | data structure procedure time generator solution
information [337]	data | result file process facility procedure feature
design [331]	implementation | application model result feature concept
operation [323]	structure representation model processor extension process
memory [318]	storage | store structure device number | data process cost
matrix [316]	table data function set | value grammar equation structure
value [300]	set error result approximation parameter | number type
storage [299]	memory | store | cost set structure file time table
error [288]	value | result data time performance solution number table
implementation [281]	design application | example feature | result data
list [277]	file | number analysis program definition | technique
file [275]	information list | input model | device procedure technique
grammar [273]	production query extension relation semantics | requirement
tree [271]	table procedure file data grammar list class generator code
scheme [268]	procedure technique | approach method model | structure
example [267]	application | implementation description | procedure
approach [266]	model | scheme procedure concept technique method process
feature [264]	extension implementation | design characteristic technique
rule [258]	procedure | technique form scheme model | approach grammar
type [256]	definition structure class syntax | operation table input
form [253]	structure element | class data representation rule table
level [242]	degree application design comparison model | structure file
property [238]	result table characteristic feature model application
concept [235]	model | approach extension design characteristic language
point [235]	set class pattern object code | structure description
class [234]	type relation | form application procedure set design

CACM. Query Experiments Results

See Page 105

CACM

	base	DOC	SEXT	stem	fam	S+fam	S+f+stem
	PRECISION						
Recall: 10	0.438	0.389	0.381	0.451	0.466	0.415	0.403
Recall: 20	0.347	0.341	0.303	0.359	0.343	0.335	0.296
Recall: 30	0.288	0.262	0.237	0.292	0.279	0.238	0.231
Recall: 40	0.251	0.203	0.196	0.246	0.230	0.183	0.188
Recall: 50	0.199	0.145	0.161	0.189	0.197	0.161	0.161
Recall: 60	0.173	0.133	0.133	0.172	0.173	0.142	0.141
Recall: 70	0.146	0.108	0.113	0.140	0.144	0.116	0.107
Recall: 80	0.118	0.091	0.091	0.119	0.115	0.094	0.089
Recall: 90	0.075	0.062	0.061	0.067	0.072	0.063	0.059
Average	0.226	0.193	0.186	0.226	0.224	0.194	0.186
Better	---	17	6	21	11	9	17
Same	---	3	8	7	16	5	3
Worse	---	32	38	24	15	38	32
	RECALL						
At 5 docs:	0.25	0.20	0.22	0.23	0.25	0.23	0.21
At 10 docs:	0.22	0.18	0.18	0.23	0.21	0.19	0.18
At 15 docs:	0.19	0.16	0.16	0.20	0.19	0.17	0.16
At 20 docs:	0.17	0.15	0.15	0.18	0.17	0.15	0.14
At 25 docs:	0.16	0.13	0.13	0.17	0.16	0.14	0.14
Better at 15	---	10	6	11	7	7	10
Same at 15	---	24	27	33	36	25	22
Worse at 15	---	18	19	8	9	20	20

CACM --- BEST IMPROVEMENTS (*see page 105*)

Base Query	*Augmented Query*	*change*
security local network network operate system distribute system	security **secure** local network network operate **operating operation** system **algorithm method procedure technique process** distribute system **algorithm method procedure technique process**	0.259 to 0.531
fast algorithm context-free language recognition parse	fast algorithm **method technique** context-free language **algorithm application model problem procedure technique process** recognition **recognizer** parse **derivation parser**	0.060 to 0.265
result parallel complexity theory pram uniform circuit	result **analysis data method model procedure program set solution technique process programming** parallel **parallelism** complexity theory pram uniform circuit	0.036 to 0.143
parallel language computation	parallel **parallelism** language **algorithm application model problem procedure technique process** computation **calculation complex computational**	0.326 to 0.357

CACM --- WORST RESULTS

Base Query	*Augmented Query*	*change*
distribute compute structure algorithm	distribute compute structure **algorithm analysis form model number operation representation technique type operate** algorithm **method technique**	0.383 to 0.234
setl high level language	setl high level language **algorithm application model problem procedure technique process**	0.348 to 0.137
portable operate system	portable **portability** operate **operating operation** system **algorithm method procedure technique process**	0.487 to 0.272
computer science principle data structure numerical generate optimization linear program algorithm khachian russian ellipsoidal algorithm complexity algorithm	computer **data** science principle data **computer information procedure result technique process** structure **algorithm analysis form model number operation representation technique type operate** numerical generate **generating** optimization **optimum** linear program **function method procedure process result time process procedure programming** algorithm **method technique** khachian russian ellipsoidal algorithm **method technique** complexity	0.442 to 0.140

CACM. Semantic Clusters.

See Page 126

implementation *as* application	language program
implementation *as* design	application
implementation *as* example	application design result
implementation *as* feature	design
information *as* data	result technique
information *as* input	file
information *as* process	data model procedure technique
information *as* result	technique
input *as* device	file
input *as* file	information
input *as* part	form
input *as* specification	description
language *as* algorithm	method
language *as* application	problem
language *as* concept	model
language *as* model	structure algorithm method
language *as* set	problem
language *as* technique	algorithm method
length *as* size	number
length *as* space	size allocation
mechanism *as* capability	facility
mechanism *as* facility	device
memory *as* store	storage device
method *as* algorithm	program system
method *as* model	algorithm result
method *as* paper	system
method *as* problem	algorithm
method *as* procedure	technique algorithm problem program result time
method *as* result	technique program time
method *as* technique	algorithm system time
method *as* time	algorithm program
model *as* algorithm	method
model *as* approach	concept
model *as* concept	language approach operation
model *as* language	algorithm
model *as* operation	process structure
model *as* process	procedure structure
model *as* result	method
model *as* structure	language
need *as* difficulty	improvement
number *as* data	problem technique function result program
number *as* function	time structure program data
number *as* set	problem function result
number *as* time	program
operation *as* concept	model feature
operation *as* function	structure data
operation *as* model	structure
operation *as* process	model structure data
operation *as* processor	process

CACM. Semantic Clusters. Two-word Terms.

See Page 145

information-retrieval *as* data-base	data-structure decision-table
information-retrieval *as* data-structure	computer-system decision-table
information-retrieval *as* retrieval-system	data-base
information-system *as* data-base	data-structure information-retrieval
information-system *as* data-item	decision-rule file-organization
information-system *as* decision-rule	storage-requirement
input-data *as* core-memory	core-storage input-output
input-data *as* core-storage	input-output
input-data *as* disk-file	total-number
input-data *as* input-output	core-storage
input-data *as* machine-time	core-storage
input-data *as* storage-device	computer-time storage-requirement
input-data *as* total-number	disk-file
input-output *as* core-memory	core-storage
input-output *as* fortran-subroutine	core-memory
input-output *as* input-data	core-memory core-storage storage-device
level-language *as* data-base	execution-time data-structure
level-language *as* data-type	data-structure
level-language *as* execution-time	data-structure
level-language *as* object-program	source-program run-time
level-language *as* software-system	storage-allocation
level-language *as* source-program	object-program
list-processor *as* garbage-collection	list-structure data-type
list-processor *as* list-cell	garbage-collection
list-processor *as* memory-location	list-structure
list-processor *as* reference-count	list-structure
list-structure *as* data-structure	decision-table
list-structure *as* data-type	data-structure decision-table
list-structure *as* garbage-collection	data-type
list-structure *as* list-processor	garbage-collection memory-location
list-structure *as* memory-location	list-processor
list-structure *as* reference-count	list-processor
log-n *as* average-time	binary-tree
log-n *as* binary-tree	average-time poin-t avl-tree
log-n *as* list-structure	data-structure
machine-language *as* character-string	source-language
machine-language *as* fortran-program	assembly-language
machine-language *as* source-language	assembly-language
machine-language *as* user-program	assembly-language tree-structure core-memory
magic-square *as* function-minimization	gamma-function
magic-square *as* matrix-inversion	gamma-function
magic-square *as* romberg-integration	matrix-inversion function-minimization
matrix-inversion *as* bessel-function	square-root
matrix-inversion *as* magic-square	gamma-function
matrix-inversion *as* romberg-integration	magic-square bessel-function
memory-requirement *as* memory-system	program-behavior garbage-collection
memory-requirement *as* program-behavior	garbage-collection memory-system simulation-model
memory-requirement *as* response-time	storage-requirement system-performance
memory-requirement *as* simulation-model	program-behavior

CACM. First-Pass Thesaurus. *See Page 131.*

algorithm :: [1248 contexts, frequency rank 4] CACM *Relat.* program, method; system; structure, model, problem, procedure, language, time, technique. *Vbs.* present, describe, use, give, require, develop, base, show, know, compare, sort, schedule.

analysis :: [422 contexts, frequency rank 20] CACM *Relat.* control. *Vbs.* use, present, give, apply, perform, employ, program, carry.

application :: [355 contexts, frequency rank 22] CACM *Relat.* program, problem, language; design, implementation, example. *Vbs.* discuss, describe, use, illustrate, give, present, make, process, program, extend, consider, automate. *Exp.* computer application (cf. system design, resource allocation), application program (cf. internal representation, file system).

code :: [238 contexts, frequency rank 34] CACM *Relat.* table; way, word, generator, string. *Vbs.* use, develop, thread, produce, give, generate, compile.

computer :: [612 contexts, frequency rank 11] CACM *Relat.* result, technique, number; processor, user, procedure, process, data. *Vbs.* use, program, aid, base, share, write, implement, operate, generate, store, simulate, process. *Exp.* computer system (cf. data structure, execution time), computer program (cf. decision table, execution time), computer science (cf. information system, data structure), computer time (cf. storage requirement, storage space), computer network (cf. data file, storage allocation), computer technology (cf. information system, data base), computer language (cf. language feature, data representation), computer application (cf. system design, resource allocation), computer resource (cf. turnaround time, computer application), computer method (cf. hash table, differential equation).

data :: [527 contexts, frequency rank 14] CACM *Relat.* function, computer, result; number, problem, program, technique; procedure, process, information. *Vbs.* use, process, present, obtain, structure, store, share, provide, analyze, set, define, base. *Exp.* data structure (cf. execution time, decision table), data type (cf. data structure, data representation), data base (cf. file organization, information system), data item (cf. decision rule, file system), input data (cf. machine time, core storage), data representation (cf. control structure, data type), data file (cf. computer network, source language), data collection (cf. information retrieval, retrieval property).

design :: [297 contexts, frequency rank 25] CACM *Relat.* application; model, result; scheme, approach, concept, example, feature, implementation. *Vbs.* describe, use, present, discuss, aid. *Exp.* system design (cf. present system, computer application), design principle (cf. list cell, file system), computer design (cf. symbol manipulation, computer application).

equation :: [268 contexts, frequency rank 29] CACM *Relat.* function; formula, grammar, expression. *Vbs.* derive, solve, use, define, involve, give.

error :: [237 contexts, frequency rank 35] CACM *Relat.* value; . *Vbs.* bound, use, program, detect, bind, correct, associate, accumulate. *Exp.* truncation error (cf. computation time, step size), roundoff error (cf. object code, memory system), error detection (cf. system design, data item).

example :: [259 contexts, frequency rank 31] CACM *Relat.* design, implementation; result, application; characteristic, range, study, description. *Vbs.* give, discuss, use, present, illustrate, detail, describe, demonstrate, provide, process, apply.

function :: [516 contexts, frequency rank 15] CACM *Relat.* set, structure, result, data; number, time, program; process, operation, formula. *Vbs.* use, define, give, compute, make, integrate, derive, associate, specify, represent, produce, perform. *Exp.* gamma function (cf. matrix inversion, magic square), precedence function (cf. precedence relation, precedence grammar), bessel function (cf. matrix inversion, square root), function minimization (cf. gamma function, romberg integration).

6.8 CISI

Name	:	CISI
Size	:	1.3 megabyte
Documents	:	1460 (Average = 140 words)
Words	:	204K
Unique words	:	12K
Source	:	IR testbed (ftp'ed from ftp.cs.cornell)
Description	:	Information Science abstracts
Queries	:	112 (Average = 88 words)

Sample Text:

- The present study is a history of the DEWEY Decimal Classification. The first edition of the DDC was published in 1876, the eighteenth edition in 1971, and future editions will continue to appear as needed. In spite of the DDC's long and healthy life, however, its full story has never been told. There have been biographies of Dewey that briefly describe his system, but this is the first attempt to provide a detailed history of the work that more than any other has spurred the growth of librarianship in this country and abroad.
- As important for staff members' individual development as was the apprenticeship in administration, perhaps the most significant attitude one acquired while working for Guy was engendered by his insistence that librarians must be interested in and knowledgeable about the content of the materials with which they dealt. His love of literature, his respect for scholarship, his admiration for good writing and reading were manifested in many ways, but most notably in his admonition that, though we were primarily a research library, we must constantly keep in mind our obligation to collect contemporary poetry, fiction and belles-letters. It was primarily up to the library staff, he felt, to be responsible for these as well as for "general" books which crossed . . .

Sample Queries :

- What problems and concerns are there in making up descriptive titles? What difficulties are involved in automatically retrieving articles from approximate titles? What is the usual relevance of the content of articles to their titles?
- What are some of the theories and practices in computer translating of texts from one national language to another? How can machine translating compete with traditional methods of translating in comprehending nuances of meaning in languages of different structures?
- There are presently fifty to one hundred technical journals being published. On the average, two new journals appear every day. In the many journals published, one to two million articles appear every year. What attempts have been made to cope with this amount of scientific and technical publication in terms of analysis, control, storage, and retrieval?

CISI (1300K) : SEXTANT results, 50 most frequent words

word [*Contexts*]	*Groups of closest words. (See page 50)*
system [1655]	service | method | technique problem program study
library [1180]	program | study number | need librarian method paper cost
information [838]	data | document result literature | study technique problem
service [646]	system | method program work | journal study need user data
problem [627]	method system aspect study | information result type concept
study [616]	research analysis | work data survey information problem
method [569]	technique | model result system problem | approach service
data [538]	information | result | analysis study method | record
result [537]	data | method analysis information paper | number study
analysis [485]	study | data result information | theory model type
term [467]	document result type information data problem technique
number [457]	result paper information list cost | time research library
need [450]	requirement interest | problem service | organization
book [448]	journal paper | material document study article | work
paper [426]	article | book result journal report study | literature
journal [375]	book | document paper | research information data service
retrieval [366]	search | index catalog technology access application source
program [358]	service system library | information method procedure
search [358]	retrieval | information index file access list concept method
literature [356]	journal | information paper | article document material
index [355]	classification list document result retrieval catalog search
technique [354]	method | procedure | approach application system model
research [349]	study work | organization information number journal analysis
document [347]	information journal | book literature result publication
science [339]	work theory study | scientist technique characteristic
development [337]	study problem application analysis research information work
work [332]	research | service paper number activity document | concept
base [330]	file | tape record | kind search retrieval center thesaurus
theory [316]	model | analysis aspect method technique | structure study
process [308]	problem method technique factor study procedure model work
language [302]	model thesaurus method | rule term system procedure theory
structure [301]	aspect relationship type model | information function
field [295]	area | application discipline problem method literature
cost [272]	number function time result | data problem information value
collection [270]	purpose | value type literature information title publication
model [267]	theory method | procedure aspect technique approach analysis
type [258]	kind characteristic structure problem analysis set purpose
form [257]	type number | catalog service information method program
area [254]	field | discipline technique literature journal | aspect
user [254]	service technique type | number search book organization
concept [251]	approach | characteristic technique problem aspect method
approach [245]	concept method technique aspect | model type | analysis
scientist [240]	research group librarian paper science literature time
group [237]	number need service | category scientist purpose difference
measure [236]	criterion method technique result data model approach | level
year [236]	work literature number time purpose study volume paper need
article [225]	paper | report | literature book | study work document
title [225]	number | citation literature document value paper item
librarian [224]	library people scientist service journal comparison | user
catalog [222]	catalogue | index access retrieval form list | program

CISI. Query Experiments Results

See Page 105

CISI

	base	DOC	SEXT	stem	fam	S+fam	S+f+stem
	P R E C I S I O N						
Recall: 10	0.316	0.317	0.282	0.336	0.314	0.303	0.309
Recall: 20	0.232	0.224	0.215	0.240	0.247	0.225	0.237
Recall: 30	0.181	0.168	0.169	0.191	0.204	0.189	0.192
Recall: 40	0.155	0.134	0.137	0.163	0.169	0.156	0.154
Recall: 50	0.132	0.114	0.118	0.134	0.141	0.130	0.128
Recall: 60	0.110	0.091	0.094	0.113	0.112	0.098	0.106
Recall: 70	0.084	0.078	0.073	0.083	0.090	0.079	0.080
Recall: 80	0.065	0.064	0.060	0.065	0.067	0.063	0.062
Recall: 90	0.036	0.049	0.040	0.040	0.040	0.040	0.043
Average	0.146	0.138	0.132	0.152	0.154	0.143	0.146
Better	---	31	21	37	36	30	31
Same	---	1	5	3	9	3	3
Worse	---	44	50	36	31	42	42
	R E C A L L						
At 5 docs:	0.24	0.25	0.21	0.24	0.23	0.23	0.23
At 10 docs:	0.24	0.23	0.22	0.25	0.24	0.23	0.23
At 15 docs:	0.22	0.20	0.21	0.23	0.23	0.22	0.21
At 20 docs:	0.21	0.20	0.20	0.22	0.22	0.21	0.20
At 25 docs:	0.20	0.19	0.19	0.21	0.20	0.20	0.20
Better at 15	---	23	13	17	23	23	23
Same at 15	---	22	39	46	38	29	30
Worse at 15	---	31	24	13	15	24	33

CISI --- BEST IMPROVEMENTS (*see page 105*)

Base Query	*Augmented Query*	*change*
model cluster search base classification document cluster suggest efficient file organization document retrieval system possible information document effectiveness system ability distinguish relevant non-relevant document improve probabilistic model cluster search base query classification model test retrieval experiment indicate effective heuristic cluster search cluster search base model effective full search document compare query efficiency implementation model discuss	model **method procedure theory process theoretical** cluster **clustering** search base classification **class** document **information journal information-selection journal-evaluation** cluster **clustering** suggest efficient file organization document **information journal information-selection journal-evaluation** retrieval system **method service** possible information **data information-selection** document **information journal information-selection journal-evaluation** effectiveness **performance quality** system **method service** ability distinguish relevant **relevance** non-relevant **non-relevant-documents** document **information journal information-selection journal-evaluation** improve probabilistic model **method procedure theory process theoretical** cluster **clustering** search base query classification **class** model **method procedure theory process theoretical** test retrieval experiment **experimental** indicate effective heuristic cluster **clustering** search cluster **clustering** search base model **method procedure theory process theoretical** effective full search document **information journal information-selection journal-evaluation** compare **comparable comparison** query efficiency **performance** implementation model **method procedure theory process theoretical** discuss	0.077 to 0.198

CISI --- BEST IMPROVEMENTS, cont. (*see page 105*)

Base Query	*Augmented Query*	*change*
author cocitation literature measure intellectual structure show map area science case information science author unit analysis cocitations pair author variable indicate distance analysis assume author cite close raw data cocitation count draw online social scisearch social science index period gthe result map show identifiable author group school information science location group respect degree centrality peripherality author group proximity author group group boundary border author connect area research position author respect map axis arbitrary set span divergent group order aid interpretation cocitation analysis author offer technique contribute understanding intellectual structure science possible area extent area rely serial publication technique establish author document effective unit analyze specialty	author cocitation literature **information information-selection** measure intellectual **public** structure **aspect** show map area **field** science **scientific scientist** case information **data information-selection** science **scientific scientist** author unit analysis **data result study** cocitations pair author variable indicate distance analysis **data result study** assume author cite **citation** close raw data **information result information-selection** cocitation count draw online social **society** scisearch social **society** science **scientific scientist** index **index-language-devices indexing** period gthe result **analysis data information method information-selection** map show identifiable author group school information **data information-selection** science **scientific scientist** location group respect degree centrality peripherality author group proximity **orientation** author group group boundary border author connect area **field** research **study work** position author respect map axis arbitrary set span divergent group order aid **assistance** interpretation cocitation analysis **data result study** author offer technique **method procedure process** contribute understanding intellectual **public** structure **aspect** science **scientific scientist** possible area **field** extent area **field** rely serial **serial-publications serials** publication technique **method procedure process** establish author document **information journal information-selection journal-evaluation** effective unit analyze specialty	0.141 to 0.253

CISI --- BEST IMPROVEMENTS, cont. (*see page 105*)

Base Query	*Augmented Query*	*change*
threshold value boolean retrieval system appear analyze recent development problem process document retrieval system query express boolean expression purpose continue analysis show concept threshold value resolve problem inherent relevance weight explore possible evaluation mechanism retrieval document base fuzzy-set-theoretic	threshold value boolean **boolean-functions** retrieval system **method service** appear analyze recent development problem **method** process **procedure** document **information journal information-selection journal-evaluation** retrieval system **method service** query express boolean **boolean-functions** expression purpose continue analysis **data result study** show concept **approach** threshold value resolve problem **method** inherent relevance **relevant** weight explore possible evaluation mechanism retrieval document **information journal information-selection journal-evaluation** base fuzzy-set-theoretic	0.025 to 0.136
computerize information system field chemistry	computerize **computerized-library-catalog-file** information **data information-selection** system **method service** field **area** chemistry **chemical chemical-technology**	0.233 to 0.341

CISI --- WORST RESULTS

Base Query	*Augmented Query*	*change*
information dissemination journal periodical	information **data information-selection** dissemination journal **book document journal-evaluation** periodical	0.321 to 0.231
information science possible	information **data informatin-selection** science **scientific scientist** possible	0.249 to 0.089
characteristic medlars medical literature analysis retrieval system project undertake national library medicine index medical journal index system index medicus major component medlars project major operate detail	characteristic **aspect** medlars medical **medicine** literature **information information-selection** analysis **data result study** retrieval system **method service** project undertake national library **librarian** medicine **medical** index **index-language-devices indexing** medical **medicine** journal **book document journal-evaluation** index **index-language-devices indexing** system **method service** index **index-language-devices indexing** medicus major component **computer-readable** medlars project major operate **operational** detail	0.442 to 0.231

CISI. Semantic Clusters.

See Page 126

index *as* catalog	retrieval
index *as* classification	list
index *as* list	classification catalog search
index *as* retrieval	search
index *as* search	retrieval
information *as* analysis	data result study
information *as* data	result study method
information *as* document	literature
information *as* literature	document
information *as* method	problem
information *as* number	result
information *as* problem	study
information *as* result	data study problem method
information *as* study	problem
information *as* technique	result problem method
institution *as* discipline	community
interest *as* area	literature
interest *as* need	problem
interest *as* requirement	need
item *as* material	document journal book literature
item *as* source	journal
journal *as* data	information
journal *as* document	book information
journal *as* material	book document
journal *as* paper	book
kind *as* application	purpose
kind *as* characteristic	type aspect
kind *as* content	nature
kind *as* nature	characteristic
kind *as* part	purpose
list *as* catalog	index
list *as* classification	index
list *as* file	search
list *as* index	search
list *as* number	information
list *as* search	index
literature *as* article	paper material book
literature *as* document	information book
literature *as* material	article document book
literature *as* paper	study book
literature *as* publication	document
literature *as* work	paper
material *as* article	book literature
material *as* document	book literature information journal
material *as* journal	book
material *as* literature	book information document
method *as* approach	technique model
method *as* data	result information
method *as* model	technique result

CISI. Semantic Clusters. Two-word Terms.

See Page 145

index-language *as* classification-system	subject-area document-collection
index-language *as* document-collection	index-term document-retrieval
index-language *as* ir-system	relevance-judgement
index-language *as* relevance-judgement	ir-system
index-language *as* retrieval-performance	index-term
index-term *as* document-collection	document-retrieval
index-term *as* index-language	document-collection retrieval-performance
index-term *as* information-retrieval	data-base
index-term *as* library-system	natural-language retrieval-system
index-term *as* natural-language	retrieval-system data-base information-retrieval
index-term *as* retrieval-system	data-base information-retrieval
information-center *as* information-service	information-science information-system
information-center *as* library-resource	library-system
information-explosion *as* citation-analysis	literature-search
information-explosion *as* computer-system	social-system
information-flow *as* communication-system	information-transfer present-paper
information-flow *as* information-need	information-service information-science
information-flow *as* social-science	information-need information-service
information-need *as* information-flow	social-science
information-need *as* information-retrieval	data-base
information-need *as* information-science	information-system information-retrieval
information-need *as* information-service	information-system retrieval-system
information-need *as* information-system	information-science information-retrieval
information-need *as* library-service	information-system information-science
information-need *as* library-system	information-system retrieval-system
information-need *as* retrieval-system	information-system information-science
information-need *as* social-science	information-science information-service
information-network *as* cable-television	computer-technology
information-network *as* computer-technology	cable-television
information-retrieval *as* index-term	data-base retrieval-system
information-retrieval *as* information-need	information-system data-base retrieval-system
information-retrieval *as* information-science	information-system data-base
information-retrieval *as* information-service	information-system data-base retrieval-system
information-retrieval *as* information-storage	information-system retrieval-system
information-retrieval *as* information-system	data-base information-science
information-retrieval *as* natural-language	data-base retrieval-system information-science
information-retrieval *as* retrieval-system	information-system data-base
information-science *as* information-need	information-system information-retrieval
information-science *as* information-retrieval	data-base
information-science *as* information-service	information-system information-retrieval
information-science *as* information-storage	information-system information-retrieval
information-science *as* information-system	information-retrieval data-base
information-science *as* library-service	information-system information-service
information-science *as* natural-language	information-retrieval retrieval-system
information-science *as* retrieval-system	information-system information-retrieval data-base
information-science *as* social-science	information-service data-base information-need
information-scientist *as* information-flow	information-need social-science
information-service *as* information-need	information-system information-science
information-service *as* information-retrieval	data-base

CISI. First-Pass Thesaurus. *See Page 131.*

analysis :: [485 contexts, frequency rank 10] CISI *Relat.* result, data; problem, information, study; approach, research, type, model, theory. *Vbs.* use, show, make, suggest, reveal, base, give, provide, follow, detail, describe, consider. *Exp.* system analysis (cf. computer technology, library system), cost analysis (cf. computer program, past decade), content analysis (cf. search request, document retrieval), text analysis (cf. document retrieval, retrieval effectiveness), subject analysis (cf. data element, document description), citation analysis (cf. literature search, citation index).

area :: [254 contexts, frequency rank 38] CISI *Relat.* field; literature; interest, discipline. *Vbs.* give, publish, increase, specialize, represent, present, make, indicate. *Exp.* subject area (cf. index language, information scientist), problem area (cf. growth rate, periodical literature), research area (cf. social organization, social structure).

base :: [330 contexts, frequency rank 27] CISI *Relat.* retrieval, search; access, thesaurus, center, kind, record, tape, file. *Vbs.* use, develop.

book :: [448 contexts, frequency rank 14] CISI *Relat.* paper; study, data; literature, article, document, material, journal. *Vbs.* use, intend, write, publish, make, find, deal, present, design, describe, consider, concern. *Exp.* book catalog (cf. card catalog, catalog card), book selection (cf. subject specialist, marc tape), present book (cf. reference work, library work).

collection :: [270 contexts, frequency rank 34] CISI *Relat.* type; list, source, purpose. *Vbs.* use, organize, grow, give. *Exp.* document collection (cf. index term, classification system), library collection (cf. library user, library staff), core collection (cf. user requirement, library resource).

cost :: [272 contexts, frequency rank 33] CISI *Relat.* process; information, result, number; performance, value, time, function. *Vbs.* use, give, relate, operate, minimize, make, determine. *Exp.* total cost (cf. document description, system design), cost factor (cf. subject matter, research library), cost analysis (cf. computer program, past decade), cost accounting (cf. research project, user need).

data :: [538 contexts, frequency rank 8] CISI *Relat.* analysis, problem, method, study, result; information; paper, journal, record. *Vbs.* collect, use, obtain, provide, present, analyze, process, hold, gather, contain, base, record. *Exp.* data base (cf. information retrieval, retrieval system), data element (cf. search term, subject analysis), survey data (cf. graduate student, content analysis).

development :: [337 contexts, frequency rank 25] CISI *Relat.* application. *Vbs.* lead, use, make, describe, consider.

document :: [347 contexts, frequency rank 23] CISI *Relat.* literature, journal; term, book, information; material, reference, publication. *Vbs.* retrieve, index, use, select, infer, classify, analyze, want, obtain, find, contain. *Exp.* document collection (cf. index term, classification system), document description (cf. decision theory, total cost), document retrieval (cf. text analysis, content analysis), document classification (cf. decision theory, classification system).

field :: [295 contexts, frequency rank 32] CISI *Relat.* literature; discipline, area. *Vbs.* relate, make, give, provide, apply.

form :: [257 contexts, frequency rank 37] CISI *Relat.* type; catalog. *Vbs.* take, use, represent, present, modify.

6.9 CRAN

Name	:	CRAN
Size	:	1.6 megabyte
Documents	:	1400 (Average = 180 words)
Words	:	260K
Unique words	:	11.8K
Source	:	IR testbed (ftp'ed from ftp.cs.cornell)
Description	:	Aeronautic abstracts
Queries	:	225 (Average = 18 words)

Sample Text:

- experimental investigation of the aerodynamics of a wing in a slipstream . an experimental study of a wing in a propeller slipstream was made in order to determine the spanwise distribution of the lift increase due to slipstream at different angles of attack of the wing and at different free stream to slipstream velocity ratios . the results were intended in part as an evaluation basis for different theoretical treatments of this problem .
- the boundary layer in simple shear flow past a flat plate . the boundary-layer equations are presented for steady incompressible flow with no pressure gradient .
- effect of uniformly distributed roughness on turbulent skin-friction drag at supersonic speeds . an experimental program was carried out in the 18-in. by 20-in. supersonic wind tunnel of the jet propulsion laboratory to determine the effect of uniformly distributed sand-grain roughness on the skin-friction drag of a body of revolution for the case of a turbulent boundary layer . the mach number range covered was 1.98 to 4.54, and the reynolds number varied from about 3 x 10 to 8 x 10 . some data were also obtained at a mach number of 0.70 . at speeds up to a mach number of 5 and for roughness sizes such that the quadratic resistance law holds, the compressibility effect is indirect, and the skin-friction drag is a function of only the roughness reynolds number, exactly as in the ...

Sample Queries :

- what similarity laws must be obeyed when constructing aeroelastic models of heated high speed aircraft .
- what problems of heat conduction in composite slabs have been solved so far .
- papers on flow visualization on slender conical wings .

CRAN (1600K) : SEXTANT results, 50 most frequent words

word [*Contexts*]	*Groups of closest words. (See page 50)*
flow [899]	equation distribution pressure | case solution effect
number [659]	value | velocity result coefficient | pressure ratio effect
layer [492]	region | case condition surface wall boundary-laye transfer
effect [456]	number theory pressure | flow distribution case result
theory [399]	method analysis | equation solution problem | number result
pressure [395]	flow | number temperature velocity speed | value effect
distribution [389]	coefficient | flow profile | result ratio drag solution
solution [373]	equation | method | result theory | value distribution
result [338]	method | solution number | distribution data value theory
ratio [335]	coefficient | distribution number profile | increase
equation [328]	solution | theory flow | problem result | transfer
method [317]	theory result solution | analysis value number equation
body [270]	cone | case flow | surface wing nozzle field plate aerofoil
wave [238]	disturbance | transfer effect edge distribution layer wing
transfer [233]	friction | problem region flow convection equation layer
field [220]	distribution calculation case | transfer term data section
speed [219]	mach tunnel pressure temperature stream | number velocity
condition [201]	result | layer number mach distribution range pressure data
velocity [195]	number | pressure coefficient mach value | temperature
value [193]	number | result solution velocity pressure expression
boundary [192]	mix | boundary-laye case flow region | equation pressure
wing [191]	plate | distribution number velocity body case ratio layer
coefficient [187]	measurement derivative distribution drag | number ratio
problem [171]	theory equation | case transfer analysis | number
temperature [167]	pressure property number speed | velocity conduction angle
angle [164]	temperature mach | distribution pressure drag number velocity
plate [159]	wall | wing stream fluid surface case | parallel change
surface [159]	layer | plate element distribution coefficient cone flow
case [156]	flow problem layer boundary | field consideration effect
analysis [149]	theory | investigation | study | measurement problem
shape [132]	distribution data | profile cone pressure coefficient
fluid [131]	plate friction | case liquid layer wall atmosphere boundary
profile [131]	distribution | characteristic data ratio | variation shape
heat [129]	load | problem temperature stress convection property
measurement [129]	coefficient | calculation investigation study distribution
transition [129]	characteristic thickness noise | coefficient number case
data [124]	information | result profile shape | investigation basis
investigation [123]	study | analysis work determination experiment | test
range [123]	change | condition increase analysis value limit tunnel
tunnel [119]	speed | investigation base range compressor drag | jet air
gradient [117]	density | difference component distribution | direction
region [116]	stream | layer pressure transfer | part boundary
study [110]	investigation | work determination test | analysis
thickness [103]	friction | transition interaction ratio profile equation
drag [100]	coefficient | distribution increase lift | measurement
gas [99]	smoke air type jet range flow phenomenon problem viscosity
stream [97]	region | speed | turbulence boundary-laye plate jet
calculation [96]	measurement | field result data application equation term
tube [95]	probe pipe | drag pressure modification profile technique
rate [92]	element friction | property coefficient distribution

CRAN. Query Experiments Results

See Page 105

CRAN

	base	DOC	SEXT	stem	fam	S+fam	S+f+stem
	PRECISION						
Recall: 10	0.745	0.551	0.666	0.741	0.715	0.667	0.665
Recall: 20	0.633	0.458	0.560	0.632	0.607	0.550	0.552
Recall: 30	0.525	0.392	0.459	0.529	0.513	0.450	0.455
Recall: 40	0.455	0.349	0.387	0.456	0.453	0.380	0.389
Recall: 50	0.394	0.313	0.339	0.400	0.401	0.336	0.337
Recall: 60	0.310	0.255	0.252	0.321	0.323	0.265	0.266
Recall: 70	0.230	0.196	0.183	0.235	0.237	0.187	0.182
Recall: 80	0.180	0.166	0.148	0.185	0.189	0.149	0.145
Recall: 90	0.121	0.118	0.100	0.123	0.128	0.103	0.100
Average	0.399	0.311	0.344	0.402	0.396	0.343	0.343
Better	---	65	44	87	70	49	59
Same	---	4	17	39	51	10	11
Worse	---	156	164	99	104	166	155
	RECALL						
At 5 docs:	0.37	0.29	0.32	0.37	0.36	0.31	0.31
At 10 docs:	0.27	0.23	0.24	0.28	0.28	0.24	0.24
At 15 docs:	0.22	0.20	0.19	0.22	0.23	0.19	0.19
At 20 docs:	0.19	0.17	0.16	0.19	0.19	0.16	0.16
At 25 docs:	0.16	0.14	0.14	0.16	0.17	0.14	0.14
Better at 15	---	52	25	26	36	31	35
Same at 15	---	87	121	176	163	107	106
Worse at 15	---	86	79	23	26	87	84

CRAN --- BEST IMPROVEMENTS (*see page 105*)

Base Query	*Augmented Query*	*change*
internal slip flow heat transfer study	internal slip **slip-flow** flow **distribution effect pressure distance present** heat **characteristic heat-transfer pressure temperature variation velocity heat-transfer heat present vary** transfer **friction layer transition** study **investigation**	0.488 to 0.807
experimental technique shell vibration	experimental **experiment** technique **data method procedure** shell **plate** vibration **frequency vibratory**	0.473 to 0.783
information buckle sandwich sphere	information **accuracy prediction** buckle **instability stability buckling** sandwich **sandwich-type** sphere **spherical**	0.375 to 0.679
dome contour minimize discontinuity stress closure cylindrical pressure vessel	dome contour minimize **minimal** discontinuity **discontinuous** stress **load loading** closure cylindrical **cylinder** pressure **flow number temperature velocity present** vessel	0.389 to 0.611

CRAN --- WORST RESULTS

Base Query	*Augmented Query*	*change*
subsonic transonic flutter data measure langley transonic dynamics tunnel compare similar data obtain facility	subsonic transonic **transonic-bump** flutter data **calculation measurement result** measure langley **stabilizer** transonic **transonic-bump** dynamics **wind-tunnel wind** tunnel compare **swirl yield** similar data **calculation measurement result** obtain facility	0.612 to 0.271
information pertain effect slight rarefaction boundary layer flow slip effect	information **accuracy prediction** pertain effect **coefficient condition distribution flow pressure problem result solution theory distance present theoretical** slight rarefaction boundary **distribution flow pressure region boundary-layer distance present** layer **case condition distribution number region surface transfer distance** flow **distribution effect pressure distance present** slip **slip-flow** effect **coefficient condition distribution flow pressure problem result solution theory distance present theoretical**	0.769 to 0.369
result creep buckle round tube external pressure	result **solution** creep buckle **instability stability buckling** round **rounded-entrance** tube external pressure **flow number temperature velocity present**	0.610 to 0.180
information pertain boundary layer slender body revolution continuum flow transverse curvature effect	information **accuracy prediction** pertain boundary **distribution flow pressure region boundary-layer distance present** layer **case condition distribution number region surface transfer distance** slender body **cone surface wing conical** revolution **independent** continuum **continuum-type** flow **distribution effect pressure distance present** transverse **compression compressive** curvature effect **coefficient condition distribution flow pressure problem result solution theory distance present theoretical**	0.704 to 0.265

CRAN. Semantic Clusters.

See Page 126

increase *as* change	variation
increase *as* decrease	reduction change
increase *as* rate	coefficient
increase *as* reduction	change
increase *as* variation	ratio
influence *as* reduction	increase
information *as* accuracy	prediction
information *as* review	description
injection *as* diffusion	flux
instability *as* buckle	stability
instability *as* mode	stability
integration *as* treatment	consideration
interaction *as* presence	pattern
interaction *as* separation	thickness transition transfer
investigation *as* analysis	result problem
investigation *as* calculation	analysis measurement data result
investigation *as* data	measurement result
investigation *as* experiment	study data
investigation *as* measurement	data
investigation *as* study	calculation analysis measurement data
investigation *as* work	study experiment
jet *as* stream	gas
layer *as* condition	distribution number
layer *as* profile	distribution
layer *as* rate	temperature
length *as* dimension	size radius
length *as* geometry	diameter
length *as* radius	diameter
length *as* size	radius
lift *as* derivative	characteristic
lift *as* force	pressure stress
lift *as* load	force pressure characteristic stress
line *as* direction	location plane
line *as* plane	direction location
load *as* characteristic	coefficient
load *as* force	stress pressure field
load *as* lift	pressure force
load *as* stress	temperature
location *as* direction	plane
location *as* movement	position
location *as* plane	direction
location *as* position	distance
mach *as* velocity	speed temperature pressure
magnitude *as* determination	expression
magnitude *as* term	variation
mean *as* detail	extension
measurement *as* calculation	data investigation result analysis
measurement *as* data	investigation result
measurement *as* investigation	data analysis

CRAN. Semantic Clusters. Two-word Terms.

See Page 145

indicial-lift *as* moment-function	total-lift
indicial-lift *as* spanwise-distribution	total-lift
indicial-lift *as* total-lift	lift-coefficient
influence-coefficient *as* control-point	twist-distribution
influence-coefficient *as* twist-distribution	unswept-wing lift-distribution
influence-coefficient *as* unswept-wing	lift-distribution taper-ratio
initial-imperfection *as* deformation-theory	stability-theory test-data
initial-imperfection *as* internal-pressure	external-pressure
initial-imperfection *as* plastic-range	deformation-theory test-data
initial-imperfection *as* stability-theory	deformation-theory
integral-equation *as* downwash-distribution	kernel-function
integral-equation *as* mach-line	kernel-function
integral-method *as* body-shape	displacement-thickness
integral-method *as* boundary-layer-equation	prandtl-number velocity-profile
integral-method *as* displacement-thickness	body-shape von-karman
integral-method *as* karman-pohlhausen-method	boundary-layer-equation
interference-effect *as* base-pressure	wind-tunnel
internal-pressure *as* elastic-core	external-pressure ring-stiffened-cylinder
internal-pressure *as* external-pressure	von-karman
internal-pressure *as* initial-imperfection	external-pressure
internal-pressure *as* ring-stiffened-cylinder	elastic-core
internal-pressure *as* stress-distribution	wall-thickness external-pressure
internal-pressure *as* wall-thickness	stress-distribution
jet-effect *as* divergence-angle	total-pressure
jet-effect *as* free-stream-flow	rocket-jet
jet-effect *as* jet-exhaust	rocket-jet flat-plate-wing
jet-effect *as* rocket-jet	total-pressure
jet-effect *as* total-pressure	pressure-ratio
jet-engine *as* jet-effect	rocket-jet divergence-angle
jet-engine *as* jet-noise	aircraft-structure
jet-flow *as* external-flow	pitot-tube
jet-noise *as* external-load	aircraft-structure fatigue-failure
jet-noise *as* fatigue-failure	aircraft-structure stress-level
jet-noise *as* jet-engine	aircraft-structure
jet-noise *as* present-state	fatigue-failure
jet-noise *as* pressure-fluctuation	aircraft-structure stress-level
jet-noise *as* stress-level	aircraft-structure fatigue-life pressure-field
jet-pressure *as* base-diameter	divergence-angle
jet-pressure *as* boattail-angle	base-diameter divergence-angle
jet-pressure *as* divergence-angle	total-pressure
jet-pressure *as* jet-off-condition	base-diameter divergence-angle
jet-pressure *as* jet-velocity	stream-velocity
kernel-function *as* downwash-distribution	integral-equation lift-distribution
kernel-function *as* lift-distribution	delta-wing
kernel-function *as* mach-line	integral-equation
kernel-function *as* twist-distribution	lift-distribution load-distribution
kernel-function *as* velocity-potential	lift-distribution
lift-coefficient *as* delta-wing	aspect-ratio
lift-coefficient *as* lift-curve-slope	lift-drag-ratio cross-sectional-area
lift-distribution *as* downwash-distribution	kernel-function velocity-potential
lift-distribution *as* taper-ratio	unswept-wing

CRAN. First-Pass Thesaurus. *See Page 131.*

analysis :: [700 contexts, frequency rank 28] CRAN *Relat.* investigation; method, problem, theory, equation, result, solution; case, study, calculation. *Vbs.* present, make, use, give, buckle, base, consider, simplify, show, linearize, extend, restrict. *Exp.* present analysis (cf. slender body theory, prandtl number), flutter analysis (cf. non uniform motion, boundary layer characteristic), stress analysis (cf. design problem, pressure vessel).

angle :: [750 contexts, frequency rank 24] CRAN *Relat.* ratio, number; . *Vbs.* combine, vary, increase, modulate, give, tabulate, range, make, exist, turn, show, provide. *Exp.* zero angle (cf. newtonian theory, surface pressure), yaw angle (cf. stagnation line, heat transfer distribution), sweep angle (cf. taper ratio, flutter speed ratio), entry angle (cf. closed form solution, corridor depth), semivertex angle (cf. shock wave shape, wall thickness), semiapex angle (cf. pressure measurement, force coefficient), initial angle (cf. pitch rate, airfoil shape), divergence angle (cf. base diameter, jet pressure), boattail angle (cf. profile shape, stability characteristic).

body :: [1085 contexts, frequency rank 14] CRAN *Relat.* wing; flow, layer; cylinder, plate, surface, cone. *Vbs.* point, nose, incline, consider, apply, present, heat, cool, blunt, vary, make, give. *Exp.* body shape (cf. shock shape, displacement thickness), body diameter (cf. jet exit diameter, total pressure), body surface (cf. pressure coefficient, momentum equation), body force (cf. temperature difference, heat addition), body length (cf. thickness distribution, angle of attack range), body theory (cf. tail surface, cruciform wing).

boundary :: [848 contexts, frequency rank 20] CRAN *Relat.* flow; surface, field, region. *Vbs.* separate, cool, heat, give, move, consist, characterise. *Exp.* boundary layer (cf. mach number, reynolds number), boundary layer equation (cf. navier stokes equation, integral method), boundary layer flow (cf. skin friction, profile shape), boundary layer thickness (cf. displacement thickness, boundary layer flow), boundary layer transition (cf. california institute, roughness element), boundary layer theory (cf. temperature gradient, yaw angle), boundary layer separation (cf. section shape, rocket jet), boundary layer problem (cf. transverse curvature, entropy layer), boundary layer characteristic (cf. present method, separation point), boundary layer growth (cf. external flow, flow region). *Fam.* boundary-layer.

case :: [605 contexts, frequency rank 36] CRAN *Relat.* characteristic, analysis, field; distribution, flow, number, layer, problem; . *Vbs.* consider, apply, give, present, obtain, investigate, discuss, solve, show, extend, examine, limit.

characteristic :: [616 contexts, frequency rank 35] CRAN *Relat.* case, data; pressure, problem, distribution, coefficient; load, test, derivative. *Vbs.* determine, calculate, predict, obtain, investigate, associate, linearize, use, show, load, derive, alter. *Exp.* flutter characteristic (cf. center of gravity location, sweptback wing), stability characteristic (cf. free flight measurement, boattail angle), control characteristic (cf. force test investigation, scale model), performance characteristic (cf. flow compressor, air jet).

coefficient :: [877 contexts, frequency rank 19] CRAN *Relat.* value; ratio, number; derivative, function, parameter, characteristic, velocity, temperature, rate. *Vbs.* measure, obtain, give, calculate, vary, determine, show, find, thrust, indicate, increase, compare. *Exp.* lift coefficient (cf. lift drag ratio, delta wing), drag coefficient (cf. nose shape, fineness ratio), pressure coefficient (cf. zero angle, heat transfer coefficient), moment coefficient (cf. force coefficient, total lift), discharge coefficient (cf. momentum equation, jet pressure), influence coefficient (cf. control point, unswept wing), force coefficient (cf. moment coefficient, taper ratio), accommodation coefficient (cf. relaxation time, skin temperature).

6.10 HARVARD

Name	:	HARVARD
Size	:	3.9 Megabyte
Documents	:	8523 (Average = 78 words)
Words	:	665 K
Unique words	:	50.6 K
Source	:	Groliers
Description	:	Sentences containing an **institution** hyponym in WordNet

Sentences were extracted from Groliers containing any of the following strings:

harvard institution establishment charity religion faith church vicariate vicarship school educational academy honorary society foundation bank commercial bank orphanage orphan asylum penal institution constitution establishment formation initiation founding foundation institution origination set up creation instauration colonization settlement

Sample Text:

- In the Roman Catholic and Anglican churches an abbey is a monastery , usually belonging to the Benedictine or Cistercian order , governed by an abbot (for communities of men) or an abbess (for communities of women) . An abbey is normally an independent institution . In Britain the term abbey is also used for such churches as WESTMINSTER ABBEY or such country houses as Woburn Abbey , which formerly belonged to monastic institutions .
- Later it was applied to the head of a religious house following a monastic rule , such as the rule of St . Abbots became important figures in MONASTICISM and church government during the Middle Ages .
- After studying at the Pennsylvania Academy of the Fine Arts , he worked as a painter and an illustrator , primarily for Harper's , and illustrated works by such authors as Herrick , Goldsmith , and Shakespeare . In 1878 , Abbey settled in England , and in 1898 he was elected to the Royal Academy of Arts .

HARVARD (3900K) : SEXTANT results, 50 most frequent words

word [*Contexts*]	*Groups of closest words. (See page 50)*
school [5580]	institution | church | settlement university work religion
church [4051]	school | institution settlement religion | constitution
institution [2077]	school system | university education settlement program
settlement [1961]	institution church | school system constitution city group
bank [1462]	government state settlement | church school constitution law
religion [1452]	religious faith institution tradition | art church culture
constitution [1391]	government | law institution | system education power
system [1292]	institution | program government constitution | power
formation [1276]	creation establishment | system church settlement religion
work [1028]	school institution group century program education study
foundation [968]	institution program development | center part constitution
government [907]	constitution | institution power education system law
education [889]	training | institution college student government | study
religious [889]	religion | institution society form state law building power
university [831]	institution | college school | academy city education
group [826]	institution form leader movement organization | education
establishment [818]	creation institution center system education program
power [791]	authority | right control government | institution
creation [785]	establishment formation | constitution system development
state [784]	law | country education system city religious group power
program [772]	institution | system college | development activity reform
law [698]	constitution state government authority institution
year [681]	century education government time student | number college
center [669]	development institution city part education establishment
art [655]	architecture | music tradition society | religion style
form [647]	group structure religion religious tradition language style
century [597]	year | number constitution religion work education group
member [588]	group year constitution number power government state
part [574]	center role area | year institution number time religion law
movement [518]	group leader development program education organization
city [506]	center town | settlement state area establishment building
faith [499]	religion | idea belief tradition doctrine law | teaching
study [493]	development education | training program activity history
academy [478]	education year college organization institution university
right [475]	freedom power | authority status | government control law
number [474]	education century year group variety state student building
role [462]	part influence position center structure power figure
music [461]	art architecture literature | drama tradition style
student [457]	child | education college year | program teacher course
life [452]	history tradition society development center education
time [436]	year period history | part life art number education
style [428]	art architecture form music | tradition painter structure
area [425]	center country part city community development land | life
tradition [423]	art religion belief | practice idea life form history
building [421]	structure development religious city center art work number
society [419]	culture art | religious institution religion organization
development [410]	study center activity | history growth reform change program
college [403]	education university | institution volume student program
order [395]	community life belief form institution society | center
authority [394]	power | control right law government | reform function

HARVARD. Semantic Clusters.

See Page 126

idea *as* belief	tradition faith practice
idea *as* concept	view theory
idea *as* doctrine	view faith belief
idea *as* practice	tradition
idea *as* principle	theory belief
idea *as* view	doctrine theory
individual *as* person	child
industry *as* firm	company
influence *as* control	power
influence *as* power	religious
institution *as* center	program
institution *as* church	school
institution *as* constitution	settlement church
institution *as* education	government
institution *as* government	system constitution
institution *as* program	system
institution *as* settlement	school church
institution *as* system	settlement constitution
institution *as* university	school
instruction *as* course	training education
instruction *as* training	education study
interest *as* activity	study
issue *as* controversy	struggle
issue *as* matter	problem
issue *as* question	matter problem
knowledge *as* skill	training
language *as* culture	art
language *as* history	tradition
language *as* literature	art architecture culture history
language *as* scholar	literature
law *as* authority	government right
law *as* constitution	institution
law *as* government	constitution institution
law *as* practice	education
law *as* right	government
leader *as* movement	group
legislature *as* delegate	representative
legislature *as* parliament	congress
life *as* center	institution
life *as* development	center
life *as* history	tradition development
life *as* organization	society
literature *as* architecture	music art history
literature *as* culture	language music art
literature *as* drama	architecture poetry
literature *as* history	language
literature *as* music	art
literature *as* paint	architecture
literature *as* scholar	language
loan *as* asset	mortgage
loan *as* assistance	aid
loan *as* credit	fund
loan *as* enterprise	fund
loan *as* mortgage	asset

HARVARD. Semantic Clusters. Two-word Terms.

See Page 145

income-tax *as* property-tax	state-law money-supply
income-tax *as* state-court	state-law
income-tax *as* state-government	school-system
income-tax *as* state-law	state-court
income-tax *as* state-right	state-chartered-bank
initiation-rite *as* church-building	religious-ritual religious-organization
interest-rate *as* bank-account	loan-association credit-union
interest-rate *as* business-firm	loan-association
interest-rate *as* credit-union	loan-association federal-reserve-bank
interest-rate *as* discount-rate	federal-reserve-bank money-supply
interest-rate *as* federal-reserve-bank	loan-association money-supply
interest-rate *as* loan-association	money-supply
interest-rate *as* member-bank	loan-association credit-union
interest-rate *as* money-supply	loan-association
interest-rate *as* mortgage-loan	loan-association credit-union
internal-improvement *as* executive-power	de-gaulle
internal-improvement *as* state-right	state-bank income-tax executive-power
land-grant-institution *as* business-administration	graduate-degree graduate-program
land-grant-institution *as* graduate-degree	state-university graduate-program
land-grant-institution *as* graduate-program	state-university graduate-degree
land-grant-institution *as* grant-bachelor	graduate-degree graduate-program
land-grant-institution *as* home-economics	graduate-degree graduate-program
land-grant-institution *as* land-grant-school	state-university
land-grant-institution *as* social-work	state-university land-grant-school
land-grant-institution *as* state-school	state-university graduate-degree
land-grant-institution *as* state-university	land-grant-school
land-grant-school *as* graduate-program	state-university land-grant-institution
land-grant-school *as* grant-bachelor	land-grant-institution graduate-program
land-grant-school *as* land-grant-institution	state-university
land-grant-school *as* library-science	state-university grant-bachelor
land-grant-school *as* social-work	state-university land-grant-institution
land-grant-school *as* state-institution	state-university law-school
law-school *as* grammar-school	art-school
law-school *as* land-grant-school	state-university
law-school *as* state-institution	land-grant-school state-university
left-bank *as* capital-city	right-bank east-bank north-bank
left-bank *as* east-bank	west-bank
left-bank *as* north-bank	right-bank east-bank
left-bank *as* religious-center	right-bank east-bank century-ad
left-bank *as* right-bank	east-bank west-bank religious-center
left-bank *as* river-bank	north-bank
left-bank *as* roman-settlement	capital-city
loan-association *as* bank-account	credit-union interest-rate
loan-association *as* business-firm	interest-rate pension-fund
loan-association *as* credit-union	interest-rate pension-fund business-firm
loan-association *as* federal-reserve-bank	interest-rate money-supply
loan-association *as* government-agency	mortgage-loan credit-union pension-fund
loan-association *as* member-bank	mortgage-loan credit-union interest-rate
loan-association *as* money-supply	interest-rate

HARVARD. First-Pass Thesaurus. *See Page 131.*

academy :: [478 contexts, frequency rank 33] HARVARD *Relat.* school, institution, university, education, year; company, college. *Vbs.* establish, found, attend, study, paint, serve, graduate, enter, continue, teach, receive, form. *Exp.* art academy (cf. art school, research center), petersburg academy (cf. school level, grant bachelor).

art :: [655 contexts, frequency rank 24] HARVARD *Relat.* government, religion; painter, culture, literature, society, style, tradition, music, architecture. *Vbs.* study, teach, patronize, find, serve, own, influence, engineer, distinguish, contribute. *Exp.* art school (cf. plaster cast, art museum), art museum (cf. symphony orchestra, art school), art academy (cf. art school, research center), art institution (cf. francisco de, dance school), religious art (cf. religious work, religious theme), visual art (cf. new york school, religious theme), art form (cf. initiation rite, folk music), art education (cf. school diploma, best actor).

bank :: [1462 contexts, frequency rank 5] HARVARD *Relat.* school, church, settlement; group, state. *Vbs.* locate, situate, lie, establish, hold, nationalize, find, allow, require, compete, charter, use. *Exp.* west bank (cf. east bank, sq mi), west bank territory (cf. sq km, sq mi), east bank (cf. west bank, left bank), left bank (cf. right bank, east bank), right bank (cf. left bank, north bank), river bank (cf. natural harbor, hudson river school), north bank (cf. capital city, right bank), blood bank (cf. school system, religious organization), stream bank (cf. soil formation, rock formation), south bank (cf. sq mi, sq km).

center :: [669 contexts, frequency rank 23] HARVARD *Relat.* program; establishment, institution; role, area, part, life, development. *Vbs.* locate, use, remain, note, bank, operate, manufacture, make. *Exp.* religious center (cf. century ad, east bank), research center (cf. research institution, library science), trade center (cf. trade route, lutheran church).

century :: [597 contexts, frequency rank 26] HARVARD *Relat.* year; work, constitution, government, education; beginning, number. *Vbs.* begin, date, found, build, continue, use, lead, develop, bring, rebuild, grow, survive.

church :: [4051 contexts, frequency rank 2] HARVARD *Relat.* school; work, bank, constitution, religion, settlement, institution. *Vbs.* establish, build, use, begin, reform, found, belong, separate, remain, ordain, serve, join. *Exp.* church music (cf. chamber music, public school), orthodox church (cf. roman catholic church, religious order), anglican church (cf. religious freedom, presbyterian church), gothic church (cf. town hall, side aisle), roman church (cf. universal church, feast day), parish church (cf. public building, byzantine church), church architecture (cf. side aisle, byzantine church), byzantine church (cf. church architecture, parish church), romanesque church (cf. gothic cathedral, abbey church), church member (cf. religious denomination, church organization).

city :: [506 contexts, frequency rank 30] HARVARD *Relat.* establishment, state, institution, settlement; community, building, area, town. *Vbs.* locate, found, wall, lie, situate, name, learn, build, note, lead, fortify, contain.

constitution :: [1391 contexts, frequency rank 7] HARVARD *Relat.* system; settlement, church, institution; amendment, century, power, law, government. *Vbs.* adopt, provide, write, approve, draft, govern, ratify, accord, suspend, give, establish, make. *Exp.* state constitution (cf. state legislature, school district), present constitution (cf. one party state, executive power).

creation :: [785 contexts, frequency rank 18] HARVARD *Relat.* establishment; work, foundation, formation, system; . *Vbs.* lead, result, encourage, call, advocate, involve, give, work, support, preside, come, bring.

6.11 JFK

Name	:	JFK
Size	:	3.2 megabyte
Documents	:	331 (Average = 1020 words)
Words	:	342 K
Unique words	:	14.8 K
Source	:	Articles on JFK assissination conspiracy theories (ftp U. Michigan)
Description	:	news articles and book extracts

Sample Text:

- Henry Kissinger did not relinquish the CIA-oriented job of National Security Advisor when he became Secretary of State. This is no doubt an unauthorized and perhaps illegal use of this position because the law requires that the president have a National Security Advisor. By his very duties this advisor performs functions that are in direct conflict with those of the Secretary of State. Since the mid-50s the Special Group or Forty Committee has become a power unto itself. The State Department has thousands of career people who are responsible for the Foreign Policy of the United States to the Forty Committee's five men. They approve items that have much greater impact on world events than the State Department.
- The FBI has overstepped their bounds in using various tactics in interviewing me. I didn't shoot John Kennedy. I didn't even know Gov. John Connally had been shot. I don't own a rifle. I didn't tell Buell Wesley Frazier anything about bringing back some curtain rods. My wife lives with Mrs. Ruth Paine.
- The shooting occurred as the presidential limousine cruised down Elm Street toward the underpass. One of the major conclusions of the commission is that the shots ''were fired from the sixth floor window at the southeast corner of the Texas School Book Depository'' (R18), a book warehouse located on the northwest corner of Elm and Houston. (Oswald was employed in this building.) Several factors influenced this conclusion. The report first calls upon the witnesses who indicated in some way that the shots originated from this source. It refers to two spectators who claimed to see ''a rifle being fired" from the depository window, two others who ''saw a rifle in this window immediately after the assassination," and ''three employees of the depository, observing the parade from the fifth floor," who ''heard the shots fired from the floor immediately above them" (R61).

JFK (3300K) : SEXTANT results, 50 most frequent words

word [*Contexts*]	*Groups of closest words. (See page 50)*
oswald [1106]	time man | commission report bullet | president cia people
evidence [974]	fact report time commission | man rifle conclusion
man [726]	oswald | people time | evidence member report witness shot
time [690]	oswald | man report evidence commission day | cia year
report [676]	commission oswald evidence story | time testimony
assassination [611]	murder | oswald evidence assassin commission part report
people [594]	man | president oswald report group witness member | agent
shot [591]	bullet | rifle president evidence jfk assassin position time
bullet [568]	shot oswald | ammunition evidence commission man | weapon
rifle [568]	weapon evidence shot fact | gun man window cia time floor
cia [487]	group oswald warren-commission time committee evidence
investigation [451]	witness staff evidence work member conspiracy case | report
story [443]	report testimony article | warren-commission position
commission [435]	report oswald warren-commission | evidence time committee
wound [424]	fragment body shot | president portion rifle wrist floor
member [395]	man investigation people report | witness oswald time cia
film [390]	evidence picture photograph photo committee fact report
president [384]	oswald jfk people shot nixon | kennedy time man commission
testimony [364]	story report statement | evidence witness information
case [338]	murder conspiracy investigation picture evidence
committee [327]	warren-commission group commission | cia nixon man
day [325]	time oswald | year man shot witness week president people
fact [288]	evidence | rifle oswald truth information film | question
group [276]	cia committee people | agent investigation member man team
position [275]	story shot location time window information report control
conclusion [266]	evidence finding theory | oswald report warren-commission
conspiracy [265]	case operation investigation truth commission cia fbi murder
fragment [265]	piece | composition wound residue shot | number ray type
work [264]	investigation staff | time support report people man
car [260]	window shot lot assassin entrance limousine man position
part [257]	point assassination support evidence | event time
office [252]	window investigation officer people bill man | department
witness [251]	investigation testimony man people | member time story
question [250]	fact time piece evidence story doubt people number effort cia
window [247]	front | building | gunman assassin top knoll rifle
book [243]	article report | story fact cia information time
power [243]	people government cia support number man | authority agency
assassin [241]	window shot gunman assassination witness involvement
article [238]	story issue book | report letter warren-commission review
information [238]	material report document evidence testimony fact file
way [231]	man conspiracy | cia number place effort work evidence
agent [230]	involvement activity weapon people group friend member
media [223]	policy | committee staff press congress government
photo [222]	photograph picture | film evidence man ft | information
control [218]	support staff conspiracy position member | effort
warren-commission [215]	commission committee | cia story | conclusion
number [211]	call fragment oswald ray name frame fact assassination series
statement [205]	testimony | story report information material fact witness
head [203]	hand neck kennedy | back motion bone visit body bullet photo
truth [203]	coverup conspiracy fact involvement investigation | scene

JFK. Semantic Clusters.

See Page 126

information *as* fact	evidence
information *as* file	material document
information *as* material	document
information *as* position	story
information *as* report	evidence
information *as* testimony	report story
interest *as* policy	media
interest *as* wud	security
investigation *as* agency	staff
investigation *as* conspiracy	case
investigation *as* staff	work
investigation *as* truth	conspiracy
investigation *as* witness	member
issue *as* article	book
jfk *as* kennedy	president
jfk *as* president	shot people
jfk *as* president-kennedy	kennedy
jfk *as* week	kennedy
kennedy *as* jfk	president
kennedy *as* president	man oswald
kennedy *as* president-kennedy	jfk
kennedy *as* week	jfk
knoll *as* drive	street
knoll *as* street	drive
letter *as* article	book
letter *as* review	article
man *as* member	people
man *as* people	oswald report
man *as* president	oswald people time shot
man *as* report	oswald time evidence
man *as* rifle	evidence shot
man *as* shot	time evidence
man *as* time	oswald evidence report
man *as* witness	people member
material *as* document	information photograph
material *as* file	information document
material *as* film	evidence
material *as* information	report
material *as* photograph	document film
member *as* agent	people
member *as* man	oswald
member *as* officer	agent
member *as* people	man report oswald
member *as* witness	man investigation people
murder *as* assassin	assassination witness
murder *as* conspiracy	case
nixon *as* committee	cia
nixon *as* president	people
office *as* department	officer

JFK. Semantic Clusters. Two-word Terms.

See Page 145

head-wound *as* ballistics-test	front-seat ballistics-evidence
head-wound *as* bullet-fragment	floor-window
head-wound *as* bullet-fragmentation	ballistics-test exit-wound neck-wound right-side
head-wound *as* exit-wound	autopsy-report
head-wound *as* front-seat	bullet-fragment
head-wound *as* neck-wound	exit-wound
head-wound *as* right-side	exit-wound autopsy-report
heart-attack *as* church-committee	staff-member
heart-attack *as* jfk-murder	umbrella-weapon
heart-attack *as* umbrella-man	umbrella-weapon
heart-attack *as* umbrella-weapon	church-committee jfk-murder umbrella-man
jfk-assassination *as* church-committee	staff-member
jfk-assassination *as* cia-people	staff-member
jfk-assassination *as* dallas-police	jim-garrison zapruder-film
jfk-assassination *as* jfk-case	news-media staff-member john-kennedy
jfk-assassination *as* jim-garrison	zapruder-film
jfk-assassination *as* john-kennedy	jim-garrison zapruder-film dallas-police
jfk-assassination *as* kennedy-assassination	jim-garrison zapruder-film dallas-police
jfk-assassination *as* news-media	john-kennedy jim-garrison zapruder-film
jfk-assassination *as* staff-member	kennedy-assassination dallas-police
jfk-case *as* gerald-ford	pcg-member
jfk-case *as* jfk-assassination	john-kennedy jim-garrison news-media
jfk-case *as* john-kennedy	jim-garrison
jfk-case *as* king-case	pcg-member gerald-ford
jfk-case *as* news-media	jfk-assassination john-kennedy jim-garrison
jfk-case *as* pcg-member	gerald-ford
jfk-case *as* staff-member	jfk-assassination
jim-garrison *as* clay-shaw	warren-report john-kennedy
jim-garrison *as* dallas-police	zapruder-film warren-report
jim-garrison *as* jfk-assassination	zapruder-film dallas-police news-media
jim-garrison *as* jfk-case	jfk-assassination news-media john-kennedy
jim-garrison *as* john-kennedy	clay-shaw zapruder-film warren-report
jim-garrison *as* kennedy-assassination	clay-shaw zapruder-film jfk-assassination
jim-garrison *as* news-media	clay-shaw zapruder-film jfk-assassination
jim-garrison *as* zapruder-film	warren-report
john-kennedy *as* cia-agent	clay-shaw
john-kennedy *as* clay-shaw	jim-garrison warren-report
john-kennedy *as* dallas-police	jim-garrison warren-report zapruder-film
john-kennedy *as* jfk-assassination	news-media jim-garrison dallas-police
john-kennedy *as* jfk-case	news-media jfk-assassination jim-garrison
john-kennedy *as* jim-garrison	warren-report zapruder-film
john-kennedy *as* news-media	jfk-assassination clay-shaw jim-garrison
john-kennedy *as* zapruder-film	warren-report
kennedy-assassination *as* cia-agent	clay-shaw john-kennedy
kennedy-assassination *as* clay-shaw	jim-garrison john-kennedy
kennedy-assassination *as* dallas-police	jim-garrison zapruder-film
kennedy-assassination *as* editorial-policy	news-media
kennedy-assassination *as* jfk-assassination	jim-garrison news-media john-kennedy
kennedy-assassination *as* jim-garrison	zapruder-film

JFK. First-Pass Thesaurus. *See Page 131.*

article :: [238 contexts, frequency rank 36] JFK *Relat.* book; report, story; chapter, review, letter, issue. *Vbs.* publish, write, appear, present.

assassin :: [241 contexts, frequency rank 35] JFK *Relat.* car, witness, window; assassination, man, shot; route, murder, gunman. *Vbs.* fire, allege, show, protect, point, exculpate. *Fam.* assassination.

assassination :: [611 contexts, frequency rank 6] JFK *Relat.* time, report; evidence, oswald; fact, part, commission, assassin, murder. *Vbs.* involve, attempt, relate, investigate, know, follow, carry, plan, commit, cover, use, solve. *Exp.* jfk assassination (cf. news media, kennedy assassination), kennedy assassination (cf. jfk assassination, jim garrison), assassination shot (cf. assassin theory, sixthfloor window), assassination conspiracy (cf. press conference, coverup effort), assassination team (cf. weapon system, committee member), king assassination (cf. king case, pcg member), assassination weapon (cf. paper sack, c2766 rifle), assassination researcher (cf. news media, harold weisberg), assassination plan (cf. umbrella man, motorcade route), assassination case (cf. jfk assassination, assassination conspiracy). *Fam.* assassin.

book :: [243 contexts, frequency rank 34] JFK *Relat.* article; time, report, story. *Vbs.* publish, write, appear, use, own, contain.

bullet :: [568 contexts, frequency rank 9] JFK *Relat.* shot; oswald, time, man, evidence; weapon, ammunition. *Vbs.* cause, fire, strike, find, hit, recover, miss, enter, show, produce, conclude, travel. *Exp.* bullet fragment (cf. front seat, secret service agent), bullet fragmentation (cf. brain tissue, head wound), miracle bullet (cf. lead core, front seat), bullet wound (cf. autopsy report, secret service agent), dumdum bullet (cf. entrance wound, exit wound), bullet theory (cf. sixthfloor gunman, c2766 rifle), bullet hit (cf. sixthfloor window, right hand), bullet fragmentation (cf. brain tissue, head wound).

car :: [260 contexts, frequency rank 28] JFK *Relat.* window, assassin; shot; house, entrance, lot. *Vbs.* park, ride, use, drive, travel, say, miss, ted, own, find. *Exp.* car lot (cf. telephone call, bullet hit), car fragment (cf. lead core, front seat).

case :: [338 contexts, frequency rank 19] JFK *Relat.* film; evidence, investigation; photograph, picture, conspiracy, murder. *Vbs.* involve, make, take, reopen, spend, prove, interest, fire. *Exp.* jfk case (cf. king case, pcg member), cartridge case (cf. physical evidence, southeast corner), king case (cf. jfk case, king assassination), mlk case (cf. jfk case, king case), assassination case (cf. jfk assassination, assassination conspiracy).

cia :: [487 contexts, frequency rank 10] JFK *Relat.* evidence, time, oswald; agency, conspiracy, pcg, commission, warren-commission, committee, group. *Vbs.* work, want, use, know, involve, give, control, investigate, own, make, come, say. *Exp.* cia agent (cf. clay shaw, marina oswald), cia man (cf. chief counsel, assassination team), cia people (cf. cia involvement, media organization), cia involvement (cf. cia people, committee staff), cia document (cf. assassination researcher, physical evidence), cia control (cf. secret team member, media control).

commission :: [435 contexts, frequency rank 13] JFK *Relat.* cia; assassination, time, evidence, oswald, report; conspiracy, fbi, committee, warren-commission. *Vbs.* tell, conclude, find, consider, make, fire, assert, know, appear, want, use, take. *Exp.* commission member (cf. staff lawyer, fbi report), commission conclusion (cf. oak cliff area, oswald left).

committee :: [327 contexts, frequency rank 20] JFK *Relat.* film; time, people, cia, commission; select-committee, nixon, group, warren-commission. *Vbs.* take, select, work, send, know, call, use, testify, make, give, continue, want. *Exp.* committee member (cf. nondisclosure agreement, staff member), committee staff (cf. subpoena power, telephone call), church committee (cf. heart attack, jfk assassination). *Fam.* commitment.

conclusion :: [266 contexts, frequency rank 25] JFK *Relat.* commission, story, evidence; source, warren-commission, theory, finding. *Vbs.* draw, reach, fire, preconceive, underlie, lead, own, indicate, dictate. *Fam.* conclude.

conspiracy :: [265 contexts, frequency rank 26] JFK *Relat.* cia, commission, investigation, case; control, murder, fbi, truth, operation. *Vbs.* prove, point, cover, take, make, kill, involve, imagine, expose, conclude, believe. *Exp.* assassination conspiracy (cf. press conference, coverup effort), conspiracy theory (cf. management policy, subpoena power), conspiracy theorist (cf. assassination team, rifle practice).

day :: [325 contexts, frequency rank 21] JFK *Relat.* man, oswald, time; week, month, year. *Vbs.* work, take, wear, spend, shoot, retract, receive, murder, hold, find, die, come.

evidence :: [974 contexts, frequency rank 2] JFK *Relat.* investigation, conclusion, shot, commission, man, time, report, film, rifle, fact. *Vbs.* suppress, indicate, present, show, know, ignore, find, use, plant, produce, make, examine. *Exp.* physical evidence (cf. cartridge case, paper sack), ballistics evidence (cf. ballistics test, head wound), tangible evidence (cf. ballistics evidence, cartridge case).

fact :: [288 contexts, frequency rank 22] JFK *Relat.* assassination, film, oswald, rifle, evidence; fbi, question, information, truth. *Vbs.* know, present, make, use, state, prove, hit, contain, check, bring, appreciate, alter.

film :: [390 contexts, frequency rank 16] JFK *Relat.* evidence; movie, fact, committee, material, photograph, photo, picture. *Vbs.* show, take, appear, reveal, make, use, view, turn, sell, prove, depict, analyze. *Exp.* zapruder film (cf. floor window, warren report), nix film (cf. picket fence, hughes film), couch film (cf. northwest corner, muchmore film), hughes film (cf. muchmore film, time span), muchmore film (cf. hughes film, head shot). *Fam.* filming.

fragment :: [265 contexts, frequency rank 26] JFK *Relat.* shot, wound; fragmentation, type, ray, number, residue, composition, piece. *Vbs.* find, scatter, reveal, locate, deposit, compare, remove, recover, miss. *Exp.* bullet fragment (cf. front seat, secret service agent), bullet fragmentation (cf. brain tissue, head wound), metal fragment (cf. secret service agent, wrist wound), car fragment (cf. lead core, front seat).

group :: [276 contexts, frequency rank 23] JFK *Relat.* committee; member, man, people, cia; system, team, agent. *Vbs.* know, think, take, decide.

information :: [238 contexts, frequency rank 36] JFK *Relat.* part, position; fact, story, testimony, evidence, report; file, document, material. *Vbs.* provide, contain, classify, receive, give, come, withhold, publish, suppress, relate, pertain, obtain.

investigation :: [451 contexts, frequency rank 11] JFK *Relat.* evidence; agency, truth, case, conspiracy, work, staff, member, witness. *Vbs.* conduct, call, own, begin, stop, want, reopen, monopolize, make, involve, forget, establish.

man :: [726 contexts, frequency rank 3] JFK *Relat.* time, report; evidence, oswald; rifle, president, shot, witness, member, people. *Vbs.* know, name, make, show, say, look, find, pick, involve, identify, use, shoot. *Exp.* cia man (cf. chief counsel, assassination team), umbrella man (cf. umbrella weapon, stemmons freeway sign), radio man (cf. stemmons freeway sign, dal tex building), top management (cf. news organization, editorial position).

member :: [395 contexts, frequency rank 15] JFK *Relat.* investigation; report, people, oswald, man; officer, agent, witness. *Vbs.* sign, use, say, make, know. *Exp.* committee member (cf. nondisclosure agreement, staff member), staff member (cf. committee member, nondisclosure agreement), pcg member (cf. jfk case, king case), commission member (cf. staff lawyer, fbi report).

office :: [252 contexts, frequency rank 30] JFK *Relat.* force, department, bill, officer. *Vbs.* come, take, enter, represent, leave. *Fam.* officer.

oswald :: [1106 contexts, frequency rank 1] JFK *Relat.* fact, day, people, president, bullet, cia, commission, report, man, time. *Vbs.* indicate, fire, know, tell, shoot, carry, say, practice, make, kill, come, use. *Exp.* marina oswald (cf. walker incident, cia agent), oswald left (cf. police lineup, southeast corner), oswald window (cf. stemmons freeway sign, right front).

part :: [257 contexts, frequency rank 29] JFK *Relat.* information; assassination; rest, event, support, point. *Vbs.* take, make, release, contain.

people :: [594 contexts, frequency rank 7] JFK *Relat.* report, time; oswald, man; power, agent, member, witness, group, president. *Vbs.* fool, know, involve, lie, continue, come, allow, think, tell, say, believe, show. *Exp.* cia people (cf. cia involvement, media organization), level people (cf. executive branch, media organization).

position :: [275 contexts, frequency rank 24] JFK *Relat.* time, shot, report, story; policy, control, information, window, location. *Vbs.* take, fix, maintain.

power :: [243 contexts, frequency rank 34] JFK *Relat.* man, people; agency, authority, support, government. *Vbs.* give, know.

president :: [384 contexts, frequency rank 17] JFK *Relat.* time, man, people, shot, oswald; nixon, kennedy, jfk. *Vbs.* kill, ask, strike, shoot, hit, fire, state, elect, tell, come, think, murder. *Fam.* president-kennedy, presidential.

question :: [250 contexts, frequency rank 32] JFK *Relat.* fact; evidence; doubt, piece. *Vbs.* ask, raise, answer, embarrass, continue, concern, come.

report :: [676 contexts, frequency rank 5] JFK *Relat.* man, time; evidence, oswald; book, people, information, testimony, story, commission. *Vbs.* say, accord, mention, fire, conclude, release, make, cite, publish, present, leave, issue. *Exp.* fbi report (cf. executive session, commission member), autopsy report (cf. bullet wound, throat wound), warren report (cf. zapruder film, harold weisberg).

rifle :: [568 contexts, frequency rank 9] JFK *Relat.* shot; time, man, evidence; floor, window, gun, fact, weapon. *Vbs.* fire, use, own, find, take, hold, carry, bring, contain, store, practice, shoot. *Exp.* rifle shot (cf. right front, dal tex building), c2766 rifle (cf. paine garage, motorcade route), rifle practice (cf. sixthfloor gunman, marina oswald).

shot :: [591 contexts, frequency rank 8] JFK *Relat.* time, rifle, bullet; man, evidence; fragment, position, assassin, jfk, president. *Vbs.* fire, hear, come, miss, indicate, strike, hit, say, call, take, prove, establish. *Exp.* assassination shot (cf. assassin theory, sixthfloor window), rifle shot (cf. right front, dal tex building), head shot (cf. zapruder frame, ballistics test). *Fam.* shoot, show.

story :: [443 contexts, frequency rank 12] JFK *Relat.* commission; oswald, time, evidence, report; witness, position, warren-commission, article, testimony. *Vbs.* tell, write, change, carry, publish, appear, pursue, make, know. *Exp.* cover story (cf. john kennedy, u2 flight), news story (cf. clay shaw trial, secret service agent).

testimony :: [364 contexts, frequency rank 18] JFK *Relat.* man, oswald, evidence, report, story; fact, information, witness, statement. *Vbs.* give, cite, take, use, hear, base, swear, support, refer, read, provide, present.

time :: [690 contexts, frequency rank 4] JFK *Relat.* report, man; evidence, oswald; shot, president, year, day, cia, commission. *Vbs.* take, make, come, spend, publish, know, hear, appear, say, require, fire, elapse. *Exp.* time span (cf. singlebullet theory, secondfloor lunchroom), time period (cf. bullet hit, stemmons freeway sign), time management (cf. management policy, jfk murder).

window :: [247 contexts, frequency rank 33] JFK *Relat.* position, assassin; rifle; location, knoll, gunman, top, building, front. *Vbs.* fire, show, come, stand, break, appear. *Exp.* floor window (cf. depository building, southeast corner), sixthfloor window (cf. southeast corner, secondfloor lunchroom), depository window (cf. sixthfloor window, depository building), oswald window (cf. stemmons freeway sign, right front).

witness :: [251 contexts, frequency rank 31] JFK *Relat.* assassin; member, people, man, story, testimony, evidence, investigation; file. *Vbs.* call, interview, use, testify, make, locate, hear, fail, bribe, ask.

work :: [264 contexts, frequency rank 27] JFK *Relat.* evidence, investigation; effort, support, record, staff. *Vbs.* begin, report, look, start, say, plan, know, involve. *Fam.* working.

wound :: [424 contexts, frequency rank 14] JFK *Relat.* shot; damage, body, fragment. *Vbs.* cause, produce, inflict, suffer, relate, receive, locate, feel. *Exp.* head wound (cf. ballistics test, front seat), exit wound (cf. entrance wound, neck wound), entrance wound (cf. exit wound, dumdum bullet), wrist wound (cf. car fragment, metal fragment), neck wound (cf. exit wound, head wound), throat wound (cf. autopsy report, entrance wound), bullet wound (cf. autopsy report, secret service agent).

6.12 MED

Name	:	MED
Size	:	1 megabyte
Documents	:	1033 (Average = 167 words)
Words	:	187K
Unique words	:	14.5K
Source	:	IR testbed (ftp'ed from ftp.cs.cornell)
Description	:	Medical abstracts
Queries	:	30 (Average = 24 words)

Sample Text:

- correlation between maternal and fetal plasma levels of glucose and free fatty acids . correlation coefficients have been determined between the levels of glucose and ffa in maternal and fetal plasma collected at delivery . significant correlations were obtained between the maternal and fetal glucose levels and the maternal and fetal ffa levels . from the size of the correlation coefficients and the slopes of regression lines it appears that the fetal plasma glucose level at delivery is very strongly dependent upon the maternal level whereas the fetal ffa level at delivery is only slightly dependent upon the maternal level .
- changes of the nucleic acid and phospholipid levels of the livers in the course of fetal and postnatal development . we have followed the evolution of dna, rna and pl in the livers of rat foeti removed between the fifteenth and the twenty-first day of gestation and of young rats newly-born or at weaning . we can observe the following facts - 1. dna concentration is 1100 ug p on the 15th day, it decreases from the 19th day until it reaches a value of 280 ug 5 days after weaning . 2. rna concentration is 1400 ug p on the 15th day and decreases to 820 during the same period .

Sample Queries :

- the crystalline lens in vertebrates, including humans.
- the relationship of blood and cerebrospinal fluid oxygen concentrations or partial pressures. a method of interest is polarography.
- radioisotopes in heart scanning. mainly used in diagnosis of pericardial effusions. also used to study tumors, heart enlargement, aneurysms and pericardial thickening. technetium, rihsa, radioactive hippurate, cholegraffin are used.

MED (187K) : SEXTANT results, 50 most frequent words

word [*Contexts*]	*Groups of closest words. (See page 50)*
cell [1156]	tissue | group effect patient study change level case
patient [883]	case | child group | treatment result study day | effect
effect [650]	change response | level action activity result increase study
study [626]	change | observation case effect patient result response
case [572]	patient | study lesion type child disease treatment result
change [549]	increase study effect | response difference | decrease
level [548]	concentration | value rate excretion effect content
acid [486]	protein | activity fraction dna increase glucose ratio value
result [446]	effect response observation patient study finding group data
child [412]	patient infant | group case subject | form woman year
activity [410]	effect concentration | increase level number response content
disease [401]	lesion | case change carcinoma | patient result type
group [397]	patient | child result difference case | subject level day
response [389]	increase effect change | result reaction rate study treatment
rate [387]	increase concentration level | response value time result
increase [385]	decrease rise change | response reduction rate | difference
hormone [365]	serum | protein antigen dna | thyroid extract effect
tissue [350]	cell | growth cancer liver tumor | resistance disease lens
treatment [341]	therapy | patient administration case response | result
concentration [339]	level content | excretion value | rate ratio metabolism
defect [338]	disturbance case malformation | regurgitation type response
rat [331]	animal mouse | dog mice level | infant kidney day rabbit
method [298]	technique | procedure test | mean result study | group
pressure [286]	flow | volume | artery obstruction rate tension | serum
growth [284]	tumor | tissue increase effect development protein response
test [284]	technique method reaction | response study therapy
tumor [260]	carcinoma growth | cancer lesion | sarcoma tissue effect
blood [258]	level | tension concentration oxygen serum plasma liver
lesion [258]	disease | case cancer tumor symptom change | manifestation
therapy [256]	treatment | administration | drug response chemotherapy
cancer [255]	carcinoma | tumor lesion tissue | disease extract
type [249]	form case | change line feature pattern group defect disease
development [248]	growth change increase incidence production | response case
reaction [245]	response test | effect increase relationship growth
factor [236]	role | mechanism difference change defect aspect treatment
period [227]	time stage | group level result course duration change rate
difference [216]	change characteristic increase | rise correlation pattern
content [212]	concentration | metabolism composition | fraction ratio
protein [212]	antigen | dna hormone growth acid analysis | concentration
culture [208]	marrow | suspension extract lung serum antigen kidney
syndrome [206]	type psychosis case lesion symptom disease result group
injection [205]	administration dose | concentration time number response
time [204]	day rate period | age serum incidence injection group month
day [203]	hr hour | month week | year time patient group rat yr
value [202]	concentration level | increase rate decrease rise content
form [198]	type | case child sign | change problem result patient
fraction [196]	content | lens concentration antigen | serum preparation
dna [193]	protein | antigen hormone | mixture fraction a-crystallin
marrow [189]	liver | spleen serum suspension age kidney culture
technique [188]	method | test procedure | analysis change data dog therapy

MED. Semantic Clusters.

See Page 126

increase *as* change	effect
increase *as* decrease	rise change difference value
increase *as* difference	change
increase *as* level	effect
increase *as* rate	response level
increase *as* reduction	decrease rise
increase *as* response	change rate effect
increase *as* rise	decrease difference
increase *as* value	response rate effect level
infant *as* female	male rabbit
infant *as* male	female rabbit
infant *as* mice	rat
infection *as* disease	case
infusion *as* administration	dose
infusion *as* dose	administration
infusion *as* irradiation	dose
infusion *as* replacement	mg
injection *as* administration	dose
injection *as* dose	administration
kidney *as* experiment	marrow
kidney *as* eye	lens bone lense
kidney *as* lens	liver
kidney *as* lense	lens eye infant
kidney *as* liver	lens marrow
lens *as* eye	lense lung kidney
lens *as* lense	eye serum kidney
lens *as* liver	lung serum kidney plasma
lens *as* plasma	serum liver
lesion *as* cancer	tumor
lesion *as* case	study
lesion *as* disease	case change
lesion *as* manifestation	symptom
lesion *as* tumor	cancer
level *as* amount	concentration excretion
level *as* concentration	rate
level *as* content	concentration
level *as* excretion	concentration
level *as* increase	rate effect
level *as* rate	increase
level *as* serum	value blood
level *as* value	concentration rate effect increase
liver *as* age	marrow
liver *as* kidney	marrow
liver *as* lens	lung plasma serum kidney
liver *as* plasma	lens serum blood
liver *as* serum	marrow blood
liver *as* spleen	marrow
lung *as* eye	lens kidney epithelium
lung *as* lens	liver serum kidney

MED. Semantic Clusters. Two-word Terms.

See Page 145

ionic-strength *as* electron-micrograph	sedimentation-coefficient
ionic-strength *as* insoluble-protein	protein-fraction m-urea lens-protein
ionic-strength *as* lens-protein	protein-fraction
ionic-strength *as* m-urea	sedimentation-coefficient insoluble-protein
ionic-strength *as* protein-fraction	amino-acid lens-protein
ionic-strength *as* sedimentation-coefficient	m-urea electron-micrograph insoluble-protein
kidney-cell *as* adult-rat	dna-synthesis folic-acid rat-kidney cell-division
kidney-cell *as* cell-division	dna-synthesis
kidney-cell *as* folic-acid	dna-synthesis
kidney-cell *as* hela-cell	human-cell tissue-culture
kidney-cell *as* human-cell	tissue-culture
kidney-cell *as* lens-epithelium	dna-synthesis lymphoid-cell
left-ventricle *as* dilution-curve	right-ventricle pressure-curve ductus-arteriosus
left-ventricle *as* ductus-arteriosus	right-ventricle
left-ventricle *as* outflow-tract	right-ventricle valve-replacement pressure-curve
left-ventricle *as* pressure-curve	right-ventricle dilution-curve ductus-arteriosus
left-ventricle *as* right-ventricle	stroke-volume
left-ventricle *as* stroke-volume	right-ventricle blood-flow carbon-dioxide
left-ventricle *as* valve-replacement	right-ventricle blood-flow stroke-volume
lens-epithelium *as* cell-population	lymphoid-cell
lens-epithelium *as* compound-lipid	lymphoid-cell
lens-epithelium *as* control-animal	kidney-weight
lens-epithelium *as* kidney-cell	lymphoid-cell dna-synthesis folic-acid
lens-protein *as* gel-filtration	protein-fraction insoluble-protein
lens-protein *as* insoluble-protein	protein-fraction gel-filtration m-urea
lens-protein *as* ionic-strength	protein-fraction
lens-protein *as* lens-regeneration	protein-fraction
lens-protein *as* m-urea	gel-filtration insoluble-protein ionic-strength
lens-protein *as* protein-component	gel-filtration
lens-protein *as* protein-fraction	gel-filtration amino-acid ionic-strength
liver-cell *as* bile-duct	type-ii
liver-cell *as* cell-hepatitis	bile-duct
lung-cancer *as* cancer-patient	radiation-therapy
lung-cancer *as* cell-carcinoma	cell-line human-lung cancer-cell
lung-cancer *as* human-lung	cell-line
lung-tissue *as* cell-carcinoma	cell-line human-lung
lung-tissue *as* cell-line	tissue-culture human-cell
lung-tissue *as* electron-microscopy	tissue-culture bone-marrow
lung-tissue *as* human-cell	cell-line tissue-culture
lung-tissue *as* human-lung	cell-line tissue-culture
lung-tissue *as* plasma-cell	bone-marrow
lung-tissue *as* tissue-culture	electron-microscopy
lupus-erythematosus *as* adult-patient	visual-agnosia
lupus-erythematosus *as* case-report	collagen-disease
lupus-erythematosus *as* collagen-disease	lupus-nephritis case-report heart-disease
lupus-erythematosus *as* heart-disease	nervous-system
lupus-erythematosus *as* lupus-nephritis	collagen-disease case-report
lupus-erythematosus *as* visual-agnosia	nervous-system
lymph-node *as* cell-carcinoma	human-lung

MED. First-Pass Thesaurus. *See Page 131.*

acid :: [486 contexts, frequency rank 8] MED *Relat.* dna, fraction, hormone, activity, protein. *Vbs.* saturate, transform, mobilize, increase, extract, esterify. *Exp.* amino acid (cf. protein synthesis, protein metabolism), tenuazonic acid (cf. tumor growth, vit d), acid synthesis (cf. control kidney, rat kidney), acid phosphatase (cf. enzyme activity, electron microscopy), acid metabolism (cf. mean concentration, body temperature), folic acid (cf. rat kidney, dna content), acid composition (cf. total lipid, blood glucose).

activity :: [410 contexts, frequency rank 11] MED *Relat.* level, effect; protein, concentration, amount, number. *Vbs.* increase, show, determine, decrease, reduce, inhibit, enhance, contain, alter. *Exp.* enzyme activity (cf. hypophysectomized rat, acid phosphatase), surface activity (cf. surface tension, inclusion body).

blood :: [258 contexts, frequency rank 27] MED *Relat.* level; liver, plasma, marrow, value, serum, oxygen, tension. *Vbs.* increase, study, make, find, estimate. *Exp.* blood pressure (cf. oxygen tension, carbon dioxide), blood flow (cf. carbon dioxide, fluid po2), blood volume (cf. stroke volume, flow rate), blood glucose (cf. newborn lamb, ffa level), peripheral blood (cf. thymus cell, bone marrow), cord blood (cf. ffa level, newborn infant), blood pool (cf. age group, blood volume), blood viscosity (cf. blood cell, stress reaction), blood stream (cf. lymphoid cell, electron microscope), blood disease (cf. adult patient, dna molecule).

cancer :: [255 contexts, frequency rank 29] MED *Relat.* lesion, tumor; tissue, disease; carcinoma. *Vbs.* advance, disseminate. *Exp.* breast cancer (cf. stage iv, cancer patient), lung cancer (cf. cell carcinoma, cell line), cancer patient (cf. breast cancer, total estrogen), cancer chemotherapy (cf. survival time, intra arterial infusion), cancer cell (cf. cell carcinoma, human cell).

case :: [572 contexts, frequency rank 5] MED *Relat.* change, study; patient; result, treatment, child, defect, type, disease, lesion. *Vbs.* present, report, occur, find, describe, study, discuss, use, observe, classify, diagnose, analyze. *Exp.* case report (cf. lupus erythematosus, intra arterial infusion), case history (cf. inclusion disease, hypophysectomized rat), index case (cf. cleft palate, nervous system).

cell :: [1156 contexts, frequency rank 1] MED *Relat.* tissue. *Vbs.* label, find, infect, contain, appear, show, nucleate, culture, transfuse, transform, observe, make. *Exp.* lymphoid cell (cf. bone marrow, thymus cell), tumor cell (cf. tissue culture, hela cell), liver cell (cf. bile duct, serum protein), cell line (cf. lung tissue, tissue culture), hela cell (cf. human cell, human lung), cell culture (cf. pleuropneumonia like organism, mycoplasma strain), cell division (cf. dna synthesis, zona glomerulosa), spleen cell (cf. lymph node, tumor cell), cell type (cf. chief cell, parathyroid gland), mast cell (cf. plasma cell, surface tension).

change :: [549 contexts, frequency rank 6] MED *Relat.* study, effect; alteration, disease, pattern, rise, decrease, difference, response, increase. *Vbs.* occur, observe, show, produce, find, result, mark, induce, associate, reveal, relate, note.

child :: [412 contexts, frequency rank 10] MED *Relat.* result, group; case, patient; reaction, year, woman, form, subject, infant. *Vbs.* disturb, show, study, observe, give, bear, report, present, match, find, diagnose, develop. *Fam.* childhood.

6.13 MERGERS

Name	:	MERGERS
Size	:	5.2 Megabytes
Documents	:	1216 (Average = 377 words)
Words	:	458 K
Unique words	:	45,500
Source	:	Wall Street Journal 89
Description	:	Text had MERGER as a keyword terms in a hand-coded index field (Mergers '89)

Sample Text: (MERGERS89)

- New England Electric System bowed out of the bidding for Public Service Co. of New Hampshire, saying that the risks were too high and the potential payoff too far in the future to justify a higher offer.

  The move leaves United Illuminating Co. and Northeast Utilities as the remaining outside bidders for PS of New Hampshire, which also has proposed an internal reorganization plan in Chapter 11 bankruptcy proceedings under which it would remain an independent company.

- Mr. Rowe also noted that political concerns also worried New England Electric. No matter who owns PS of New Hampshire, after it emerges from bankruptcy proceedings its rates will be among the highest in the nation, he said. "That attracts attention . . . it was just another one of the risk factors" that led to the company's decision to withdraw from the bidding, he added.

- R.P. Scherer Corp. said it completed the $10.2 million sale of its Southern Optical subsidiary to a group led by the unit's president, Thomas R. Sloan, and other managers. Following the acquisition of R.P. Scherer by a buy-out group led by Shearson Lehman Hutton earlier this year, the maker of gelatin capsules decided to divest itself of certain of its non-encapsulating businesses. The sale of Southern Optical is a part of the program.

- Commercial-vehicle sales in Italy rose 11.4% in February from a year earlier, to 8,848 units, according to provisional figures from the Italian Association of Auto Makers.

- MacMillan Bloedel Ltd. said it plans to redeem all of its 9%, Series J debentures outstanding April 27. . . .

MERGERS (5200K) : SEXTANT results, 50 most frequent words

word [*Contexts*]	*Groups of closest words. (See page 50)*
company [7625]	concern | group | firm | analyst bid price acquisition
share [6241]	stock | stake | year shareholder bid | sale plan board
stock [2801]	share | stake shareholder | year offer board sale value
offer [2744]	bid | proposal | plan | transaction agreement acquisition
stake [2661]	share | stock interest | price shareholder investment
business [2643]	operation | concern market | sale product | asset maker
unit [2623]	subsidiary | group sale operation share | year bank firm
sale [2609]	transaction | purchase acquisition business plan interest
bid [2504]	offer | proposal | plan acquisition | transaction
year [2350]	price month | week share group | bank plan stock market
group [2316]	company concern bank | firm year executive | interest
price [2295]	value | year | plan interest profit | month week company
analyst [2174]	group company executive investor bank | year official concern
plan [2022]	proposal offer bid | agreement transaction price | year
executive [2014]	official chairman group | spokesman management analyst board
market [2003]	business | industry year concern bank investment operation
concern [1870]	company | group business firm maker | industry operation
bank [1850]	group investor | firm year airline | market transaction
agreement [1738]	plan offer | bid transaction acquisition proposal | deal
firm [1592]	concern group company | bank investor agency | year fund
officer [1517]	chairman | director president management official board
acquisition [1504]	purchase bid transaction offer | sale merger investment
interest [1442]	stake price | concern group sale investment asset value
yesterday [1440]	friday | week | month | offer plan year today time day
operation [1382]	business | concern interest asset market industry | group
official [1343]	spokesman executive | director board offer shareholder plan
spokesman [1310]	official | executive board offer | chairman shareholder
president [1307]	chairman | director officer | official spokesman board
value [1276]	price | profit cash interest amount | earning number debt
asset [1229]	interest investment operation business | debt part
investor [1168]	bank investment buyer | shareholder firm security analyst
investment [1138]	investor | transaction acquisition interest loan asset market
transaction [1130]	deal buy-out merger | offer acquisition plan purchase sale
maker [1101]	concern | manufacturer industry producer | business group
product [1088]	business service | market system equipment | sale part year
board [1037]	shareholder director | management offer executive | bid
shareholder [1013]	holder | board | share stock investor offer | acquisition
debt [979]	loan | cash bond | cost loss amount | value fund
chairman [977]	executive | director board officer president | official
purchase [952]	acquisition transaction | sale bid offer plan | investment
week [890]	month yesterday | year friday | time day price offer bid
director [883]	board | chairman official shareholder officer president
month [877]	week | year time yesterday day | transaction price bid plan
management [865]	board executive | shareholder plan manager bank officer
part [858]	value cash asset year time cost interest control fund profit
buy-out [853]	transaction | takeover | deal purchase merger acquisition
proposal [848]	offer | bid plan | transaction financing | agreement term
control [812]	stake board asset interest ownership cash investment part stock
corporation [789]	bank official group shareholder spokesman | executive
merger [770]	transaction | acquisition deal | buy-out bid takeover

MERGERS. Semantic Clusters.

See Page 126

income *as* earning	profit loss
income *as* gain	profit loss earning revenue result
income *as* increase	profit loss gain
income *as* loss	profit
income *as* profit	loss
income *as* revenue	profit earning
increase *as* decline	gain
increase *as* drop	rise decline
increase *as* gain	profit loss income
increase *as* income	profit loss
increase *as* loss	profit
increase *as* profit	loss
increase *as* rise	decline drop
individual *as* observer	familiar
industry *as* concern	market company
industry *as* maker	concern
industry *as* operation	concern market interest
industry *as* service	system
information *as* document	statement
institution *as* airline	bank security government
institution *as* carrier	airline government
institution *as* government	airline
insurance *as* consumer	network
insurer *as* chemical	retailer
insurer *as* conglomerate	chemical
insurer *as* real-estate	retailer
interest *as* asset	operation
interest *as* concern	group
interest *as* investment	asset
interest *as* operation	concern
interest *as* price	stake
interest *as* value	price
interview *as* conference	meeting
interview *as* morning	wednesday
investment *as* asset	interest
investment *as* buyer	investor
investment *as* deal	transaction acquisition takeover
investor *as* bank	analyst
investor *as* buyer	investment partner
investor *as* firm	concern
job *as* post	position
job *as* title	post
law *as* legislation	bill
law *as* regulation	regulator rule
law *as* regulator	court agency
lawsuit *as* complaint	suit
lawsuit *as* proceeding	protection
lawsuit *as* request	suit
lawsuit *as* review	request approval

MERGERS. Semantic Clusters. Two-word Terms.

See Page 145

ima-holdings-corp. *as* hospital-chain	pritzker-family hospital-company
ima-holdings-corp. *as* hospital-company	pritzker-family hospital-chain
ima-holdings-corp. *as* initial-offer	pritzker-family
ima-holdings-corp. *as* offering-price	pritzker-family
ima-holdings-corp. *as* pritzker-family	investment-group
independent-company *as* bankruptcy-proceeding	internal-reorganization rate-increase
independent-company *as* internal-reorganization	rate-increase bankruptcy-proceeding
independent-director *as* mccaw-offer	lin-share
industry-analyst *as* fiscal-year	cash-flow
industry-official *as* bailout-bill	thrift-industry
industry-official *as* distribution-system	growth-rate
industry-official *as* growth-rate	distribution-system
industry-source *as* drug-company	prescription-drug
industry-source *as* drug-maker	prescription-drug
information-service *as* community-newspaper	news-service
information-service *as* news-service	community-newspaper
information-service *as* phone-company	community-newspaper phone-line
information-service *as* phone-line	phone-company
information-system *as* human-resource	insurance-operation
insolvent-s-ls *as* budget-deficit	thrift-regulator savings-and-loan-bailout
insolvent-s-ls *as* deposit-account	deposit-rate thrift-institution
insolvent-s-ls *as* deposit-insurance	percentage-point
insolvent-s-ls *as* deposit-rate	thrift-regulator deposit-account
insolvent-s-ls *as* savings-and-loan-bailout	insolvent-thrift thrift-regulator
insolvent-s-ls *as* thrift-institution	insolvent-thrift percentage-point
insolvent-s-ls *as* thrift-regulator	insolvent-thrift budget-deficit
insolvent-thrift *as* insolvent-s-ls	thrift-regulator savings-and-loan-bailout
insolvent-thrift *as* savings-and-loan-bailout	thrift-regulator thrift-industry
insolvent-thrift *as* thrift-industry	thrift-regulator
insolvent-thrift *as* thrift-institution	insolvent-s-ls
insolvent-thrift *as* thrift-regulator	capital-requirement thrift-industry
insurance-business *as* net-asset	sale-price
insurance-business *as* price-tag	cash-reserve
insurance-commissioner *as* anglo-french-financier	insurance-regulator insurance-unit
insurance-commissioner *as* insurance-regulator	insurance-unit sir-james-goldsmith
insurance-commissioner *as* insurance-unit	sir-james-goldsmith sir-james
insurance-commissioner *as* london-based-tobacco	insurance-regulator sir-james-goldsmith
insurance-commissioner *as* sir-james-goldsmith	sir-james takeover-bid
insurance-commissioner *as* state-regulator	london-based-tobacco
insurance-commissioner *as* takeover-rule	insurance-regulator london-based-tobacco
insurance-company *as* insurance-group	insurance-unit takeover-battle
insurance-company *as* investment-company	pension-fund
insurance-company *as* pension-fund	navigation-mixte
insurance-group *as* insurance-company	navigation-mixte
insurance-group *as* takeover-battle	navigation-mixte
insurance-operation *as* financial-services-concern	london-based-tobacco
insurance-operation *as* human-resource	information-system
insurance-operation *as* third-quarter-loss	year-earlier-period

MERGERS. First-Pass Thesaurus. *See Page 131.*

acquisition :: [1504 contexts, frequency rank 22] MERGERS *Relat.* agreement; plan, offer, sale, bid; deal, investment, merger, transaction, purchase. *Vbs.* say, complete, make, propose, expect, announce, approve, seek, plan, give, require, own. *Exp.* acquisition agreement (cf. telerate share, qintex australia), acquisition proposal (cf. new york investor, buy out proposal), dd acquisition (cf. unicorp canada corp., confidentiality agreement), hal acquisition (cf. wage increase, wage concession), acquisition price (cf. lease obligation, long term debt), acquisition activity (cf. court decision, company stock).

agreement :: [1738 contexts, frequency rank 19] MERGERS *Relat.* plan; price, bid, offer; proposal, investment, transaction, deal, acquisition. *Vbs.* reach, sign, say, announce, enter, expect, terminate, make, give, propose, complete, approve. *Exp.* merger agreement (cf. board meeting, buy out group), standstill agreement (cf. jacobs group, exchange commission filing), acquisition agreement (cf. telerate share, qintex australia), option agreement (cf. c-$53, class a share), agreement call (cf. varity share, executive committee), oral agreement (cf. executive post, warner executive), loan agreement (cf. vitro s.a., apparel maker), confidentiality agreement (cf. dd acquisition, unicorp canada corp.), sale agreement (cf. product line, board approval), credit agreement (cf. cash dividend, board approval).

analyst :: [2170 contexts, frequency rank 13] MERGERS *Relat.* executive, year, group; company; official, firm, concern, bank, investor. *Vbs.* say, expect, believe, think, speculate, make, estimate, accord, predict, give, agree, consider. *Exp.* industry analyst (cf. fiscal year, cash flow), security analyst (cf. investment concern, production facility), analyst estimate (cf. mccaw offer, columbia stock), airline analyst (cf. takeover candidate, debt leverage), london analyst (cf. soft drink business, defense contractor), bank analyst (cf. creditor bank, debt leverage).

asset :: [1230 contexts, frequency rank 30] MERGERS *Relat.* operation, investment, interest; business; control, part. *Vbs.* sell, acquire, say, relate, liquidate, secure, dispose, purchase, hold, reduce, exceed, buy. *Exp.* asset sale (cf. bridge loan, cash flow), company asset (cf. family member, cash crunch), total asset (cf. sea containers ltd., olympia york), estate asset (cf. pension fund, olympia york), gas asset (cf. c-$53, gas concern), asset value (cf. co chief executive, b.a. t holder), valuable asset (cf. debt burden, film library), partnership asset (cf. long term value, exchange offer), net asset (cf. australian cent, sale price), thrift asset (cf. take advantage, bailout bill).

bank :: [1848 contexts, frequency rank 18] MERGERS *Relat.* market, analyst; year; agency, airline, thrift, investor. *Vbs.* say, hold, make, buy, acquire, agree, sell, own, lead, take, believe, operate. *Exp.* investment banker (cf. junk bond, wall street), investment bank (cf. bridge loan, buy out group), bank debt (cf. bank loan, bank financing), bank financing (cf. labor management buy out, ual stock), bank loan (cf. ual buy out, equity investment), bank lender (cf. hooker asset, past month), bank agreement (cf. mortgage note, raise cash), bank borrowing (cf. stock offering, debt financing), bank analyst (cf. creditor bank, debt leverage), bank account (cf. core company, government agency).

bid :: [2503 contexts, frequency rank 9] MERGERS *Relat.* year, offer; share; month, purchase, acquisition, agreement, transaction, plan, proposal. *Vbs.* make, say, launch, consider, raise, reject, receive, accept, sweeten, own, succeed, revise. *Exp.* takeover bid (cf. sir james goldsmith, sir james), amax bid (cf. acquisition agreement, share capital), hoylake bid (cf. b.a. t share, cash portion), davis bid (cf. ual chairman, ual share), cash bid (cf. real estate developer, b.a. t share), socanav bid (cf. socanav inc., steinberg share), rival bidder (cf. socanav inc., rival bid), rival bid (cf. socanav inc., c-$30), merieux bid (cf. connaught share, connaught shareholder), joint bid (cf. defense contractor, labor group).

6.14 MOBYDICK

Name	:	MOBY DICK
Size	:	1 Megabyte
Words	:	244 K
Unique words	:	19.6 K
Source	:	Gutenburg project
Description	:	Herman Melville's whaling adventure

Sample Text:

- Call me Ishmael. Some years ago , never mind how long precisely , having little or no money in my purse, and nothing particular to interest me on shore, I thought I would sail about a little and see the watery part of the world. It is a way I have of driving off the spleen, and regulating the circulation. Whenever I find myself growing grim about the mouth; whenever it is a damp, drizzly November in my soul; whenever I find myself involuntarily pausing before coffin warehouses, and bringing up the rear of every funeral I meet; and especially whenever my hypos get such an upper hand of me, that it requires a strong moral principle to prevent me from deliberately stepping into the street, and methodically knocking people's hats off , then, I account it high time to get to sea as soon as I can.
- Ultimately the defendants (the crew of another ship) came up with the whale, struck, killed, seized, and finally appropriated it before the very eyes of the plaintiffs. And when those defendants were remonstrated with, their captain snapped his fingers in the plaintiffs' teeth, and assured them that by way of doxology to the deed he had done, he would now retain their line, harpoons, and boat, which had remained attached to the whale at the time of the seizure. Wherefore the plaintiffs now sued for the recovery of the value of their whale, line, harpoons, and boat. Mr. Erskine was counsel for the defendants; Lord Ellenborough was the judge. In the course of the defence, the witty Erskine went on to illustrate his position, by alluding to a recent crim. con. case, wherein a gentleman, after in vain trying to bridle his wife's viciousness, had at last abandoned her upon the seas of life; but in the course of years, repenting of that step, he instituted an action to recover possession of her.

MOBYDICK (1000K) : SEXTANT results, 50 most frequent words

word [*Contexts*]	*Groups of closest words. (See page 50)*
whale [979]	man | ahab time boat thing queequeg part hand | way captain
man [805]	ahab whale ship | hand queequeg boat captain stubb thing time
ship [535]	boat | man ahab way sea vessel | queequeg thing stubb
ahab [479]	stubb queequeg man | hand ship captain starbuck whale boat
boat [424]	ship | man way hand ahab whale line starbuck | body side
sea [366]	water ship | boat deck air time whale surface hand wave
hand [355]	arm ahab eye man queequeg face | boat leg head side
head [354]	eye hand body ship thing | side ahab line part place
thing [339]	ship man whale ahab queequeg head captain hand | sir part
way [321]	ship boat face ahab | part captain water look man queequeg
time [314]	whale man hand day ship ahab queequeg captain way sea
eye [297]	hand head | face ahab part whale way boat ship look
part [293]	end man way eye body side whale ship place hand
stubb [237]	queequeg starbuck | ahab captain sir | bildad crew man ship
water [234]	sea | air day way side oil bit | ocean ship eye
captain [230]	stubb queequeg | ahab crew man starbuck ship sir bildad
queequeg [229]	stubb | ahab captain sir bildad | flask man hand ship
sort [228]	man way part stubb eye ship sight look thing sir
side [210]	bow deck bottom hand gunwale part head face water boat
day [203]	hour night time water ship boat air | foot queequeg ahab
line [175]	rope boat end leg hand head way | part ahab whale
foot [164]	length inch mile face leg live | queequeg side part day
crew [161]	captain | stubb ship bildad starbuck queequeg man way ahab
deck [155]	side bow sea | mast_head sailor pequod board arm bulwark way
arm [151]	hand | leg iron body teeth place boat finger eye way
starbuck [150]	stubb | bildad ahab sir captain daggoo | hand peleg boat
night [146]	day ship voyage tashtego side ahab mind bildad sailor morning
body [143]	fluke head bone part boat mass leg arm chase starbuck
end [142]	part turn line side | hand harpoon block morning way edge
world [141]	eye vessel captain man sail end round field distance sea
leg [139]	face hand brow arm | body button line bone foot iron
air [137]	water sea day | eye side mouth clam stream silence character
round [134]	captain leg long boat mile hand barrow part hull look
god [132]	word ahab cook sailor captain art pequod hand captain-ahab sir
face [131]	leg hand way eye side | flask look foot die queequeg
sight [123]	die voyage turn island spout heart eye tail story view
mate [122]	harpooneer jonah stubb ahab | lad starbuck flask night word
place [122]	mast_head part way arm head latitude spot shore turn hand
pequod [120]	ship queequeg god terror passage sir crew vessel | deck flame
heart [118]	brain | sight god beat home body fact soul order hand
voyage [118]	ship | passage captain seaman vessel weather sight port night
life [115]	part voyage socket peril mast spear order bone land whaleman
harpoon [111]	iron | blacksmith | lance ship boat bildad elephant
sail [110]	sailor | carpenter mast ship blow float leg battle leviathan
soul [104]	ahab boy house pip leg headsman heart wife year thought
bildad [103]	peleg sir queequeg | captain-peleg stubb starbuck blacksmith
sailor [103]	seaman sail | queequeg captain god carpenter blacksmith ahab
word [103]	god blacksmith shark queer bildad hand command rib name sailor
sir [101]	stubb queequeg bildad em | captain-ahab flask devil
bone [100]	body skeleton | ivory leg tail part rib leviathan jacket

MOBYDICK. Semantic Clusters.

See Page 126

head *as* body	part
head *as* eye	hand part
head *as* hand	ahab
head *as* place	part
head *as* side	hand part
head *as* thing	ship
leg *as* arm	hand body
leg *as* body	arm
leg *as* bone	body
leg *as* face	hand foot
leg *as* iron	arm
line *as* boat	ahab man
line *as* end	part
line *as* head	hand
line *as* leg	hand
man *as* ahab	whale ship
man *as* boat	ahab whale ship
man *as* captain	ahab queequeg stubb
man *as* hand	ahab boat
man *as* queequeg	ahab ship hand captain stubb
man *as* stubb	ahab ship queequeg captain
man *as* thing	whale ship
man *as* time	whale
night *as* mind	bildad
part *as* eye	hand
part *as* side	hand
part *as* way	ship
pequod *as* vessel	ship
place *as* arm	hand
place *as* head	hand
place *as* part	way
queequeg *as* ahab	man ship
queequeg *as* bildad	stubb sir
queequeg *as* captain	stubb ahab man
queequeg *as* flask	sir
queequeg *as* hand	ahab man
queequeg *as* ship	man
queequeg *as* sir	stubb captain bildad flask
queequeg *as* stubb	ahab captain man ship
sail *as* carpenter	sailor
sailor *as* carpenter	seaman sail
sailor *as* queequeg	captain ahab
sailor *as* seaman	carpenter
sea *as* air	water
sea *as* time	whale
ship *as* ahab	man
ship *as* boat	man ahab
ship *as* crew	stubb
ship *as* queequeg	man ahab stubb

MOBYDICK. First-Pass Thesaurus. *See Page 131.*

ahab :: [479 contexts, frequency rank 4] MOBY *Relat.* ship; whale, man; way, starbuck, boat, hand, captain, queequeg, stubb. *Vbs.* cry, say, stand, think, turn, own, mutter, hear, come, follow, step, pause.

air :: [137 contexts, frequency rank 32] MOBY *Relat.* day, sea, water; mouth. *Vbs.* curl.

arm :: [151 contexts, frequency rank 25] MOBY *Relat.* body, leg; boat, hand; place, iron. *Vbs.* throw, toss, lean, draw, cross.

boat :: [424 contexts, frequency rank 5] MOBY *Relat.* ahab; whale, man, ship; side, body, starbuck, line, hand, way. *Vbs.* lower, pull, take, stand, make, leap, jump, hoist, break, strike, shoot, reach.

body :: [143 contexts, frequency rank 28] MOBY *Relat.* arm, leg; boat, part, head; bone, mass, fluke. *Vbs.* strip, invest.

captain :: [230 contexts, frequency rank 16] MOBY *Relat.* queequeg, stubb; hand, ship, man, ahab; bildad, sir, starbuck, crew. *Vbs.* say, stand, cry, make, tell, roar, remain, look, know, command, call. *Exp.* captain ahab (cf. chief mate, sperm whale), stranger captain (cf. ivory leg, mortal man), whaling captain (cf. whaling voyage, stranger captain).

crew :: [161 contexts, frequency rank 23] MOBY *Relat.* ship, ahab, man, queequeg, stubb, captain; . *Vbs.* come, turn, float, command.

day :: [203 contexts, frequency rank 20] MOBY *Relat.* water; time; foot, air, night, hour. *Vbs.* sail, make, say.

deck :: [155 contexts, frequency rank 24] MOBY *Relat.* sea, side; pequod, sailor, mast-head, bow. *Vbs.* hoist, pace, walk, move, mount, drag, cross, come.

end :: [142 contexts, frequency rank 29] MOBY *Relat.* line, way, part; turn. *Vbs.* come, attach.

eye :: [296 contexts, frequency rank 12] MOBY *Relat.* part, way, head, hand; whale, boat, ahab; look, face. *Vbs.* look, meet, close, fix, own, lift, behold, stand, open, make, light, grow.

face :: [131 contexts, frequency rank 35] MOBY *Relat.* leg; side, eye, way, foot, hand; look.

foot :: [164 contexts, frequency rank 22] MOBY *Relat.* day, side; leg, face, mile, inch, length. *Vbs.* start, measure, exceed, stand, sit.

god :: [132 contexts, frequency rank 34] MOBY *Relat.* pequod; art, sailor, word. *Vbs.* bless, lay.

6.15 NEJM

Name	:	NEJM
Size	:	1 Megabyte
Documents	:	42 (Average = 4380 words)
Words	:	184 K
Unique words	:	8.7 K
Source	:	New England Journal of Medecine
Description	:	full text articles about AIDS
Queries	:	None, but document similarity appreciations

Sample Text:

- We conducted a serologic survey for antibodies to human immunodeficiency virus types 1 and 2 (HIV-1 and HIV-2) and human T-cell lymphotropic virus Type I (HTLV-I) in 704 Brazilians with the acquired immunodeficiency syndrome (AIDS) or at risk for it. The study population included 70 homosexual men (11 of whom were prostitutes), 58 bisexual men (19 of whom were prostitutes), 101 female prostitutes from three socioeconomic groups, 13 wives of men with hemophilia who were seropositive for HIV-1 antibodies, and 47 blood donors with positive Venereal Disease Research Laboratory tests for syphilis, all from Rio de Janeiro; 86 female prostitutes from two rural towns in Minas Gerais; 133 patients with AIDS from Sao Paulo; and 196 men with bleeding disorders who were seropositive for HIV-1 antibodies on enzyme-linked immunosorbent assay, from Sao Paulo and Rio de Janeiro. . . .
- A total of 693,000 volunteer blood donors from the Washington, D.C., area were screened for HIV infection during the 42-month period from July 1985 through December 1988. The frequency of positive Western blot tests declined from 0.14 to 0.04 percent during that time (Fig. 1). Sixteen hundred thirty-nine donors tested reactive repeatedly on enzyme immunoassay for HIV, and 284 (17 percent) of these positive enzyme immunoassays were confirmed by Western blot analysis. From this population, 156 donors who tested positive on both enzyme immunoassay and Western blot assay, 64 who tested positive on enzyme immunoassay and negative on Western blot, and 16 who tested positive on enzyme immunoassay and whose Western blot results were indeterminate were entered into the study and followed for a median of 28 months. The number of Western blot-positive donors who entered the study represented 55 percent (156 of 284) of all the Western blot-positive donors identified. Thirty-four percent did not respond to letters requesting that they contact the blood center, and 11 percent elected not to participate. The mean interval from blood donation to initial clinic visit was six weeks.

NEJM (1000K) : SEXTANT results, 50 most frequent words

word [*Contexts*]	*Groups of closest words. (See page 50)*
patient [2004]	subject | child group infection day month donor study
infection [1448]	disease | pneumonia | patient tuberculosis | aid result
level [1033]	concentration | value | count | titer number rate
group [873]	dose treatment patient subject | day man zidovudine
test [811]	assay culture | sample study data level analysis donor
study [772]	trial data | patient result rate analysis test infection
count [744]	value level | number | dose proportion change week rate
antigen [609]	antibody | protein sample donor | sequence assay | infant
dose [568]	therapy | treatment | mg week ddi group | day zidovudine
therapy [558]	treatment | dose prophylaxis | administration week effect
rate [555]	risk | time survival result | prevalence number ratio level
effect [526]	toxicity reaction | benefit | efficacy change therapy
disease [500]	infection | pneumonia | tuberculosis condition | symptom
pneumonia [493]	disease | infection tuberculosis | death aid survival
sample [487]	specimen | culture | donor antigen | donation test cell
treatment [486]	therapy | prophylaxis dose | group administration | drug
result [466]	rate data | study month donor value infection effect finding
donor [451]	donation person | man sample | woman antigen infant subject
risk [440]	rate | incidence | prevalence mortality survival case dose
subject [426]	patient man | participant person child week | group donor
antibody [414]	antigen | assay p24 infection donor level | transmission
culture [411]	sample | specimen test examination | status antigen count
cell [389]	lymphocyte | sample | plasma blood number serum pbmc virus
reaction [380]	effect | toxicity | treatment result therapy infection
analysis [373]	study model | test evaluation method | comparison result
response [365]	effect | function benefit change | status result marker
number [359]	count proportion | rate level time ratio | concentration
mg [339]	dose | ddi hour | week zidovudine | therapy placebo day
man [335]	person | subject | donor infant child | woman mother
day [327]	month | week | time | year | dose patient hour course
month [317]	day week | time | year | infant | survival interval age
week [316]	month day | time subject dose | hour year therapy course
difference [312]	change decrease respect | effect improvement increase rate
virus [310]	hiv | hiv-1 | antigen strain viremia | cell proportion
data [299]	study | result rate finding | information month test value
titer [292]	concentration | plasma level prevalence value | mortality
value [292]	level | count | concentration | cd-4 titer change
hiv-1 [286]	hiv | virus | donor sample result htlv-i person | month
infant [283]	child woman | man month | person donor mother | subject
period [265]	interval | time duration day course week survival | month
time [259]	month day | week rate survival | interval age number
assay [255]	immunoassay enzyme_linked_immunosorbent_assay kit | test
child [248]	infant | subject man person | participant patient year
aid [240]	age pneumonia infection | hiv participant disease man week
hiv [237]	hiv-1 | virus | anti-hiv-1 aid assay tuberculosis human
factor [235]	history | marker rate person difference benefit condition
incidence [225]	risk | prevalence | development mortality frequency
syndrome [225]	immunodeficiency | pancreatitis manifestation dementia type
drug [224]	agent | zidovudine | regimen medication acyclovir treatment
hospital [222]	study group area period rate population | woman month antigen

NEJM. Semantic Clusters.

See Page 126

immunoassay *as* blot	enzyme-linked-immunosorbent-assay
immunoassay *as* eia	enzyme-linked-immunosorbent-assay
immunoassay *as* immunoblotting	enzyme-linked-immunosorbent-assay blot
immunoassay *as* indeterminate	enzyme-linked-immunosorbent-assay initial
immunoassay *as* initial	indeterminate
immunoassay *as* western	indeterminate
immunodeficiency *as* pancreatitis	syndrome
improvement *as* benefit	decrease toxicity
improvement *as* change	difference
improvement *as* decrease	benefit change increase difference
improvement *as* finding	change
improvement *as* increase	change difference
improvement *as* measure	benefit
improvement *as* toxicity	benefit
incidence *as* case	risk development diagnosis
incidence *as* development	risk case diagnosis
incidence *as* frequency	risk prevalence development mortality time
incidence *as* mortality	risk prevalence frequency
incidence *as* prevalence	risk case
incidence *as* seroprevalence	prevalence
increase *as* change	difference
increase *as* decline	decrease change reduction
increase *as* decrease	change improvement difference
increase *as* drop	decrease decline reduction
increase *as* improvement	change difference
increase *as* reduction	decrease decline
infant *as* child	man month subject
infant *as* donor	antigen
infant *as* man	donor subject
infant *as* mother	woman man person
infant *as* person	child man donor subject
infant *as* subject	donor
infant *as* woman	child man person donor mother subject
infection *as* aid	disease pneumonia
infection *as* disease	pneumonia
infection *as* pneumonia	disease
infection *as* result	study effect
infection *as* study	patient
infection *as* tuberculosis	disease pneumonia
interval *as* age	time month survival
interval *as* course	period day week
interval *as* duration	period age
interval *as* month	day week
interval *as* period	time
interval *as* survival	time month week
interval *as* time	period month day week
level *as* concentration	value titer number
level *as* number	count rate

NEJM. Semantic Clusters. Two-word Terms.

See Page 145

iga-level *as* cd-cell
iga-level *as* cd-cell-count
iga-level *as* cd-ratio
iga-level *as* core-antigen
iga-level *as* igg-level
iga-level *as* neopterin-level
iga-level *as* pokeweed-mitogen
iga-level *as* tetanus-toxoid
immunodeficiency-syndrome *as* hiv--infection
immunodeficiency-syndrome *as* hiv-disease
immunodeficiency-syndrome *as* risk-factor
incidence-rate *as* age-group
incidence-rate *as* male-to-female-ratio
incidence-rate *as* seroconversion-rate
incidence-rate *as* survey-period
initial-treatment *as* average-dose
initial-treatment *as* body-surface-area
initial-treatment *as* combination-therapy
initial-treatment *as* initial-therapy
initial-treatment *as* maintenance-therapy
initial-treatment *as* median-survival
initial-treatment *as* standard-therapy
initial-treatment *as* total-dose
inosine-pranobex *as* aids-related-complex
inosine-pranobex *as* base-line
inosine-pranobex *as* cell-count
inosine-pranobex *as* placebo-group
iq-score *as* ddi-therapy
iq-score *as* oral-administration
iq-score *as* plasma-concentration
iq-score *as* weight-gain
laboratory-test *as* blood-count
laboratory-test *as* serum-antibody
laboratory-test *as* study-participant
laboratory-test *as* treatment-center
log-rank-test *as* low-dose-group
log-rank-test *as* standard-treatment-group
log-rank-test *as* treatment-group
low-dose-group *as* hemoglobin-level
low-dose-group *as* hiv-antigen
low-dose-group *as* log-rank-test
low-dose-group *as* neutrophil-count
low-dose-group *as* standard-treatment-group
low-dose-group *as* study-medication
lymphocyte-count *as* base-line
lymphocyte-count *as* ddi-therapy
lymphocyte-count *as* hemoglobin-level
lymphocyte-count *as* leukocyte-count

tetanus-toxoid cd-cell-count
tetanus-toxoid core-antigen
tetanus-toxoid pokeweed-mitogen
cd-cell-count
cd-ratio cd-cell-count cd-cell
serum-level
tetanus-toxoid confidence-interval
cd-cell-count confidence-interval
hiv-infection
aids-related-complex
hiv-infection hiv--infection
male-to-female-ratio hiv-seroprevalence
age-group hiv-seroprevalence
age-group survey-period
age-group
amphotericin-b combination-therapy
average-dose total-dose
amphotericin-b total-dose
standard-therapy
amphotericin-b median-survival
maintenance-therapy
average-dose combination-therapy
amphotericin-b combination-therapy
plasma-viremia
placebo-group cell-count
hiv-infection
cell-count
zidovudine-therapy lymphocyte-count
plasma-concentration
oral-administration
ddi-therapy
physical-examination
serum-antigen
female-prostitute
female-prostitute study-participant
treatment-group
treatment-group
placebo-group cell-count
standard-treatment-group
treatment-group
standard-treatment-group
standard-treatment-group platelet-count
treatment-group
standard-treatment-group
cell-count
platelet-count
platelet-count neutrophil-count
hemoglobin-level

NEJM. First-Pass Thesaurus. *See Page 131.*

analysis :: [366 contexts, frequency rank 25] NEJM *Relat.* test, study; basis, difference, evaluation, method, comparison, model. *Vbs.* perform, use, blot, exclude, show, indicate, plan, accord, relate, provide, confirm, conduct. *Exp.* data analysis (cf. study entry, acyclovir group), interim analysis (cf. p value, pentamidine group).

antibody :: [400 contexts, frequency rank 22] NEJM *Relat.* level, infection, sample, antigen; reactivity, seroconversion, p24, assay. *Vbs.* test, screen, detect, measure, find, direct, determine, associate. *Exp.* hiv antibody (cf. enzyme immunoassay, blood donor), antibody response (cf. schedule b, envelope epitope), serum antibody (cf. serum antigen, reverse transcriptase), antibody okt3 (cf. rejection episode, mean number).

antigen :: [603 contexts, frequency rank 8] NEJM *Relat.* specimen, month, infant, assay, sequence, hiv-1, donor, sample, protein, antibody. *Vbs.* detect, test, use, confirm, screen, measure, assay. *Exp.* p24 antigen (cf. hiv infection, hiv 1 infection), p24 antigen assay (cf. anti hiv 1 positivity, blood unit), hiv antigen (cf. hiv antigenemia, plasma viremia), hiv antigenemia (cf. hiv antigen, cd4 lymphocyte count), serum antigen (cf. serum antibody, hiv seropositive mother), core antigen (cf. culture supernatant, blot positive donor), antigen level (cf. hiv antigen, hiv antigenemia).

cell :: [389 contexts, frequency rank 23] NEJM *Relat.* sample; virus, serum, plasma, blood, lymphocyte. *Vbs.* infect, stimulate, obtain, count, use, pack, determine, activate. *Exp.* cell count (cf. base line, treatment group), t cell (cf. neopterin level, serum level), cd8 cell (cf. cd4 cell count, iga level).

count :: [739 contexts, frequency rank 6] NEJM *Relat.* level; data, ratio, change, proportion, rate, number, value. *Vbs.* perform, increase, decrease, accord, show, obtain, mean, fall, associate, use, reveal, reach. *Exp.* cell count (cf. base line, treatment group), platelet count (cf. hemoglobin level, neutrophil count), neutrophil count (cf. hemoglobin level, platelet count), blood count (cf. platelet count, serum level), lymphocyte count (cf. hemoglobin level, platelet count), leukocyte count (cf. entry value, hemoglobin level).

culture :: [408 contexts, frequency rank 21] NEJM *Relat.* sample; antigen, test; status, examination, specimen. *Vbs.* perform, test, use, obtain, isolate, define, consider, prepare, grow, detect, confirm. *Exp.* hiv culture (cf. end point dilution culture, tissue culture infective dose), culture supernatant (cf. reverse transcriptase, plasma sample), blood culture (cf. bronchoalveolar lavage specimen, bronchoalveolar lavage fluid), sputum culture (cf. bronchoalveolar lavage specimen, amphotericin b).

data :: [300 contexts, frequency rank 34] NEJM *Relat.* count, rate, result, test, study; status, aid, finding, information. *Vbs.* suggest, show, provide, indicate, analyze, use, collect, represent, report, obtain, monitor, compare. *Exp.* data analysis (cf. study entry, acyclovir group), laboratory data (cf. maintenance therapy, hiv disease).

day :: [327 contexts, frequency rank 30] NEJM *Relat.* week, month; group, dose, patient; course, hour, child, year, time. *Vbs.* die, obtain, occur, receive, give, treat, reach, diagnose, persist, perform, mean, follow.

donor :: [437 contexts, frequency rank 19] NEJM *Relat.* subject, sample; antigen; hiv-1, population, woman, infant, man, person, donation. *Vbs.* transplant, blot, test, screen, infect, confirm, identify, find, expose, select, paid, neutralize. *Exp.* blood donor (cf. serum sample, htlv i infection), volunteer donor (cf. anti hiv 1 positivity, hiv 1 transmission), donor population (cf. total number, core antigen). *Fam.* donation.

6.16 NPL

Name	:	NPL
Size	:	3.2 Megabyte
Documents	:	11429 (Average = 42 words)
Words	:	490K
Unique words	:	XX
Source	:	IR testbed (ftp'ed from ftp.cs.cornell)
Description	:	what text is about
Queries	:	100 (Average = 10 words)

Sample Text:

- COMPACT MEMORIES HAVE FLEXIBLE CAPACITIES. A DIGITAL DATA STORAGE SYSTEM WITH CAPACITY UP TO BITS AND RANDOM AND OR SEQUENTIAL ACCESS IS DESCRIBED.
- THE BRITISH COMPUTER SOCIETY. REPORT OF A CONFERENCE HELD IN CAMBRIDGE JUNE.
- D PACKAGING REDUCES SIZE OF ELECTRONIC UNITS. GREATER COMPONENT DENSITIES ARE OBTAINABLE USING A MODULE TECHNIQUE IN WHICH MINIATURE CIRCUIT ELEMENTS ARE PLACED SIDE BY SIDE WITH ELECTRICAL CONNECTION MACE ON A THREE DIMENSIONAL BASIS BY A SPOTWELDING PROCESS.
- SWITCHING CIRCUITS USING BIDIRECTIONAL NONLINEAR IMPEDANCES. A GENERAL REVIEW OF CIRCUIT LOGIC IS DEVELOPED FOR A BIDIRECTIONAL NONLINEAR SWITCHING ELEMENT. THE DESIGN OF PNP TRANSISTOR DRIVER SRIVER STAGES IS CONSIDERED. A BINARY OCTAL DECODER CIRCUIT AND A SIMPLE BINARY FULL ADDER CIRCUIT ARE DISCUSSED AS EXAMPLES.
- THE SQUARE LOOP FERRITE CORE AS A CIRCUIT ELEMENT. THE SHAPE OF THE OUTPUT WAVEFORMS WHEN THE CORES ARE SWITCHED IS EXPLAINED BY A QUANTITATIVE THEORY WHICH TAKES INTO ACCOUNT THE RESIDUAL LOSS REASONABLE AGREEMENT WITH EXPERIMENTAL EVIDENCE IS SHOWN.

Sample Queries :

- METHODS OF APPROXIMATING THE FREQUENCY PHASE RELATIONSHIPS FOR RESISTIVE INDUCTIVE AND RESISTIVE CAPACITIVE CIRCUITS
- DIURNAL VARIATIONS OF FLUCTUATIONS IN THE EARTHS MAGNETIC FIELD
- DERIVATION OF THE COMPONENTS OF THE ELECTRICAL CONDUCTIVITY IN THE UPPER ATMOSPHERE
- TEMPERATURE INDEPENDENT METHODS FOR TUNING HIGHLY STABLE HIGH FREQUENCY OSCILLATORS

NPL (3200K) : SEXTANT results, 50 most frequent words

word [*Contexts*]	*Groups of closest words. (See page 50)*
circuit [4272]	network | characteristic | method function | effect
field [4031]	function | characteristic voltage density | analysis
amplifier [3347]	filter | network | application | element function current
frequency [3189]	value | result distribution effect | analysis density
method [2715]	theory analysis | result technique | system circuit
effect [2679]	variation | characteristic | frequency result theory
theory [2582]	method analysis | equation result characteristic | effect
result [2516]	analysis | data value measurement characteristic method
variation [2477]	change | distribution effect measurement value
wave [2368]	oscillation | distribution current system density effect
system [2255]	circuit | characteristic | function application method
measurement [1943]	observation | variation | data result characteristic
analysis [1762]	result | characteristic calculation investigation application
characteristic [1754]	analysis | variation property circuit distribution system
observation [1586]	measurement | data | result record investigation variation
function [1583]	coefficient | value curve term variation | system relation
type [1577]	design | application function effect analysis measurement
distribution [1538]	variation | value change characteristic | measurement
network [1521]	circuit | amplifier | characteristic oscillator | system
value [1480]	variation | frequency distribution | result function
equation [1453]	theory | analysis relation problem | solution condition
filter [1447]	amplifier | oscillator impedance | system unit
voltage [1409]	impedance | function field value | system component
density [1371]	intensity | velocity value | change rate coefficient energy
oscillator [1311]	generator | operation output filter network | oscillation
electron [1266]	particle | radiation | component variation ionization
application [1247]	analysis characteristic design | amplifier system technique
current [1220]	temperature | component time condition | amplitude velocity
radiation [1201]	emission | source density absorption electron energy
time [1182]	rate | current temperature ratio change power | density
design [1159]	application | operation analysis type | performance
data [1027]	observation | result measurement | variation analysis
condition [1017]	characteristic | value analysis variation parameter current
line [1010]	function resistance characteristic | frequency variation time
signal [998]	current | source system amplitude transmission oscillation
noise [941]	oscillation phenomenon | energy component pulse impedance
range [932]	band | variation value characteristic curve pulse change
pulse [930]	characteristic | response range source oscillation variation
source [916]	component radiation part signal measurement | characteristic
relation [877]	variation value characteristic analysis | dependence
emission [874]	radiation | oscillation ionization intensity source
oscillation [873]	wave oscillator | fluctuation behavior noise | emission
element [870]	device | function design unit oscillator amplifier system
component [869]	current source | characteristic coefficient distribution
technique [867]	method | application | system device analysis data
region [863]	layer | distribution variation source current | measurement
power [858]	ratio energy characteristic | impedance amplitude temperature
layer [846]	region | distribution temperature | change surface
transistor [846]	diode | rectifier | unit output element device supply
temperature [841]	current | resistance value ratio | amplitude impedance

NPL. Query Experiments Results

See Page 105

	NPL						
	base	DOC	SEXT	stem	fam	S+fam	S+f+stem
	PRECISION						
Recall: 10	0.415	..NA.	0.385	0.419	0.359	0.329	0.344
Recall: 20	0.332	..NA.	0.318	0.342	0.296	0.279	0.287
Recall: 30	0.260	..NA.	0.244	0.272	0.242	0.227	0.233
Recall: 40	0.225	..NA.	0.211	0.232	0.209	0.193	0.196
Recall: 50	0.181	..NA.	0.168	0.187	0.171	0.159	0.159
Recall: 60	0.149	..NA.	0.140	0.149	0.144	0.133	0.132
Recall: 70	0.109	..NA.	0.101	0.112	0.101	0.096	0.099
Recall: 80	0.081	..NA.	0.077	0.085	0.078	0.074	0.075
Recall: 90	0.053	..NA.	0.048	0.054	0.052	0.048	0.048
Average	0.200	..NA.	0.188	0.206	0.183	0.171	0.175
Better	---	15	7	51	25	24	34
Same	---	5	5	4	13	8	3
Worse	---	73	73	38	55	56	56
	RECALL						
At 5 docs:	0.23	0.17	0.21	0.23	0.20	0.17	0.16
At 10 docs:	0.22	0.16	0.22	0.23	0.21	0.19	0.19
At 15 docs:	0.21	0.15	0.21	0.22	0.19	0.19	0.19
At 20 docs:	0.21	0.15	0.20	0.21	0.19	0.19	0.19
At 25 docs:	0.19	0.14	0.18	0.19	0.19	0.18	0.18
Better at 15	---	20	8	18	7	10	17
Same at 15	---	26	74	67	57	51	46
Worse at 15	---	47	11	8	29	32	30

NPL --- BEST IMPROVEMENTS (*see page 105*)

Base Query	*Augmented Query*	*change*
effect solar flare absorption cosmic radio noise ionosphere	effect solar flare absorption cosmic radio **radiation** noise **feedback** ionosphere **ionospheric**	0.392 to 0.488
estimate density ionization temperature solar corona	estimate **trace** density ionization temperature solar corona **coronal**	0.439 to 0.512
determination ion masse ionosphere study back scatter radio wave	determination ion **departure molecule molecular** masse **mass** ionosphere **ionospheric** study **basis investigation** back scatter **scattering** radio **radiation** wave	0.116 to 0.184
produce minimal net logical function canonical form	produce minimal net logical **logic** function canonical form	0.279 to 0.347

NPL --- WORST RESULTS

Base Query	*Augmented Query*	*change*
supply information performance typical magnetic film memory system circuit diagram	supply **converter transistor conversion convertor transformer transisto transistron transmittance** information performance typical **typica** magnetic film memory **logic logical** system **syst** circuit diagram	0.258 to 0.065
model experiment aurora	model **distribution** experiment aurora **auroral**	0.335 to 0.135
transistor sweep generator	transistor **diode transformer transisto transistron transmittance** sweep generator	0.361 to 0.150
measurement plasma temperature arc discharge shock wave technique	measurement **mean** plasma temperature arc discharge shock **sho** wave technique	1.000 to 0.500

NPL. Semantic Clusters.

See Page 126

impedance *as* gain	resistance ratio response
impedance *as* loss	ratio
impedance *as* parameter	resistance coefficient
impedance *as* resistance	power temperature
incident *as* scatter	diffraction
incident *as* transmission	propagation
increase *as* absorption	height
increase *as* change	variation distribution
increase *as* dependence	change fluctuation
increase *as* distribution	variation
increase *as* disturbance	absorption
increase *as* fluctuation	change absorption disturbance
increase *as* intensity	height absorption
inductance *as* conductance	reactance
inductance *as* inductor	reactance reactor
inductance *as* resistor	capacitor
influence *as* dependence	change relation
input *as* amplification	gain
input *as* gain	ratio response
input *as* supply	generator control
instrument *as* apparatus	equipment spectrometer
instrument *as* probe	spectrometer
intensity *as* absorption	density height
intensity *as* activity	height
intensity *as* amplitude	height
intensity *as* increase	absorption height change
intensity *as* ionization	absorption
interpretation *as* examination	comparison
interpretation *as* explanation	agreement
interpretation *as* hypothesis	explanation
interpretation *as* possibility	explanation
investigation *as* analysis	result
investigation *as* calculation	analysis
investigation *as* determination	calculation
investigation *as* discussion	analysis
investigation *as* measurement	result
investigation *as* observation	measurement result
investigation *as* study	analysis observation
ionization *as* absorption	radiation height
ionization *as* disturbance	absorption
ionization *as* emission	radiation
ionization *as* intensity	absorption height
ionization *as* reflection	echo
ionosphere *as* record	observation
irregularity *as* inhomogeneity	perturbation
irregularity *as* movement	drift
irregularity *as* reflection	ionization drift
latitude *as* activity	height storm disturbance intensity
latitude *as* altitude	zone

NPL. Semantic Clusters. Two-word Terms.

See Page 145

image-impedance *as* ladder-filter	image-parameter tchebycheff-type
image-impedance *as* unit-step	ladder-filter
image-parameter *as* circuit-component	filter-design
image-parameter *as* design-data	ladder-filter
image-parameter *as* filter-design	circuit-component
image-parameter *as* insertion-loss	band-pass pass-filter
image-parameter *as* ladder-filter	image-impedance tchebycheff-type
image-parameter *as* tchebycheff-type	insertion-loss pass-filter
incident-wave *as* cylinder-axis	point-source
incident-wave *as* plane-surface	diffraction-field point-source
incident-wave *as* plane-wave	em-wave
incident-wave *as* surface-impedance	plane-wave
input-circuit *as* glass-tube	temperature-coefficient
input-impedance *as* frequency-response	phase-shift band-pass equivalent-circuit
input-impedance *as* junction-transistor	equivalent-circuit
input-impedance *as* output-impedance	cathode-follower
input-impedance *as* phase-shift	frequency-response band-pass
input-impedance *as* power-gain	output-impedance transistor-amplifier
input-impedance *as* transient-response	frequency-response phase-shift
input-impedance *as* transistor-amplifier	junction-transistor
input-pulse *as* monostable-multivibrator	pulse-amplifier
input-pulse *as* output-pulse	input-voltage
input-signal *as* input-voltage	output-voltage
input-signal *as* output-pulse	input-voltage
input-signal *as* sine-wave	square-wave
input-signal *as* square-wave	transistor-circuit
input-signal *as* supply-voltage	output-voltage
input-signal *as* time-constant	output-voltage
input-signal *as* transistor-circuit	junction-transistor
input-voltage *as* dc-amplifier	output-voltage feedback-loop
input-voltage *as* feedback-loop	dc-amplifier
input-voltage *as* input-signal	output-voltage
input-voltage *as* output-impedance	output-voltage input-impedance
input-voltage *as* output-pulse	input-signal pulse-width
input-voltage *as* pulse-width	rise-time
insertion-loss *as* circuit-element	pass-band band-pass transfer-function
insertion-loss *as* ladder-network	pass-band band-pass transfer-function
insertion-loss *as* pass-band	band-pass transfer-function pass-filter
insertion-loss *as* pass-filter	band-pass transfer-function
insertion-loss *as* tchebycheff-type	image-parameter pass-band pass-filter
insertion-loss *as* transfer-function	band-pass
insertion-loss *as* transmission-line	band-pass equivalent-circuit
integral-equation *as* boundary-condition	plane-wave em-wave
integral-equation *as* diffraction-problem	boundary-condition
integral-equation *as* plane-wave	em-wave
integral-equation *as* transmission-coefficient	power-series
ion-concentration *as* collision-frequency	electron-concentration
ion-concentration *as* mass-spectrometer	radio-observation

NPL. First-Pass Thesaurus. *See Page 131.*

amplifier :: [3347 contexts, frequency rank 3] NPL *Relat.* unit, element, application, network, filter. *Vbs.* use, describe, distribute, couple, give, tune, design, base, discuss, connect, ground, control. *Exp.* transistor amplifier (cf. junction transistor, input impedance), dc amplifier (cf. feedback loop, output voltage), stage amplifier (cf. output impedance, noise figure), feedback amplifier (cf. phase shift, pass filter), amplifier circuit (cf. push pull, output stage), band amplifier (cf. group delay, attenuation characteristic), pulse amplifier (cf. input pulse, ge diode), power amplifier (cf. output transformer, push pull), amplifier stage (cf. output impedance, output stage), amplifier design (cf. output impedance, smith chart). *Fam.* amplification.

analysis :: [1762 contexts, frequency rank 13] NPL *Relat.* characteristic; theory, method, result; study, relation, discussion, application, investigation, calculation. *Vbs.* give, indicate, show, base, present, make, use, detail, apply, simplify, extend, develop. *Exp.* harmonic analysis (cf. sunspot activity, sunspot minimum), network analysis (cf. network problem, state response), matrix analysis (cf. impedance converter, passive element), amplifier analysis (cf. difference equation, saturation effect), fourier analysis (cf. diurnal component, frequency multiplication), circuit analysis (cf. conductance amplifier, data system).

application :: [1247 contexts, frequency rank 27] NPL *Relat.* design; amplifier, type, method, system, analysis, characteristic; operation, property, technique. *Vbs.* discuss, describe, illustrate, indicate, give, consider, note, use, show, mention, outline, switch. *Exp.* circuit application (cf. design principle, field effect), application part (cf. transistor theory, transistor parameter).

characteristic :: [1754 contexts, frequency rank 14] NPL *Relat.* measurement, analysis; system, circuit, variation; condition, relation, distribution, application, property. *Vbs.* give, discuss, operate, determine, use, describe, derive, observe, investigate, make, calculate, obtain. *Exp.* characteristic impedance (cf. image impedance, ladder network), frequency characteristic (cf. stage amplifier, input impedance), attenuation characteristic (cf. pass filter, group delay), transfer characteristic (cf. optimum filter, transistor characteristic), phase characteristic (cf. distortion factor, amplitude characteristic), valve characteristic (cf. component value, supply voltage), characteristic equation (cf. feedback network, oscillation frequency), transmission characteristic (cf. quartz crystal, filter network), transistor characteristic (cf. external feedback, transistor amplifier), response characteristic (cf. filter section, tchebycheff type).

circuit :: [4272 contexts, frequency rank 1] NPL *Relat.* field; method, characteristic, network. *Vbs.* use, describe, tune, couple, give, switch, print, discuss, design, derive, base, apply. *Exp.* transistor circuit (cf. junction transistor, equivalent circuit), circuit diagram (cf. af amplifier, junction transistor), circuit element (cf. band pass, transfer function), circuit parameter (cf. noise figure, input signal), oscillator circuit (cf. feedback loop, frequency stability), feedback circuit (cf. cathode follower, output impedance), circuit detail (cf. pulse generator, output signal), amplifier circuit (cf. push pull, output stage), filter circuit (cf. design curve, characteristic impedance), valve circuit (cf. transistor circuit, input impedance).

condition :: [1017 contexts, frequency rank 33] NPL *Relat.* current; variation, analysis, characteristic; problem, case, parameter. *Vbs.* derive, determine, operate, discuss, satisfy, give, investigate, consider, show, indicate, establish, use. *Exp.* boundary condition (cf. diffraction problem, integral equation), stability condition (cf. passive quadripole, electron stream), load condition (cf. dc supply, fall time), initial condition (cf. state solution, boundary condition), resonance condition (cf. q factor, series resonance), limit condition (cf. temperature stabilization, anode circuit), equilibrium condition (cf. uniform plasma, plasma wave), state condition (cf. cross relaxation, output frequency).

6.17 SPORTS

Name	:	SPORTS
Size	:	6 Megabyte
Documents	:	5750, Documents are groups of extracted sentences that were in the same article (Average = 81 words)
Words	:	1.1 M
Unique words	:	88 K
Source	:	Groliers Encyclopedia
Description	:	Extracted any sentence containing one of strings below These words were under SPORT in WordNet

Any sentences containing one of the following words was extracted from *Grolier's*: acrobatics alai angling archery association athletic athletics badminton baseball basketball battledore bicycling bloodsport bobsledding boxing cast casting court cricket croquet cycling dip dive diving doubles equitation field fight fishing fisticuffs flying football game golf grappling gymnastics handball handspring handstand hockey horseback horseshoes hunt hunting jai judo jujitsu jump jumping lacrosse lawn luging mare match medal miniature motorcycling outdoor pelota play plunge polo pugilism pushball quoits racquets rassling riding roller rounders row rowing royal rugby running scuba sculling shovelboard shuffleboard shuttlecock singles skating ski skiing sledding snorkel snorkeling soccer softball split sport sports squash stickball stroke surf surfboarding surfing surfriding swim swimming tennis tetherball tobogganing track tumble tumbling volleyball water wrestling

Sample Text:

- For the island village of Saynatsalo , Aalto designed a civic center (1950-52) with intimately scaled red-brick structures of various shapes clustered around an elevated grass court that affords vistas of the surrounding lake and forests
- Aaron began playing professionally for all-black teams in Mobile , Ala
  He reached the major leagues when he was only 20 and quickly established himself as one of the game's finest players . He played for the Braves almost exclusively , first in Milwaukee (1954-65) , then in Atlanta (1966-74) . 305 , Aaron had 2 , 297 runs batted in (1st all-time) , 6 , 856 total bases (1st) , 12 , 364 at bats (2d) , 3 , 771 hits (3d) , 3 , 298 games played (3d) , and 624 doubles (8th) . Aaron was the NL's Most Valuable Player in 1957 , and the right fielder won 3 Gold Glove awards for his fielding prowess .
- This shape minimizes water resistance in the abalones' intertidal habitats . The animals respire and discharge wastes through a row of holes on one side of the shell ; old holes fill up and new ones appear as the animals age .

SPORTS (6000K) : SEXTANT results, 50 most frequent words

word [*Contexts*]	*Groups of closest words. (See page 50)*
water [5199]	field | part form court material | game region number air
field [3648]	surface water | system area game court | development part
surface [3502]	field | form part skin system | structure number type work
court [3114]	supreme-court field water play game player year | government
play [2407]	work | game form | player court year field style number
role [2376]	part | number development form group field work activity
game [1599]	sport player | play number | form work year group field
system [1553]	field form area surface structure | game group development
form [1245]	type surface structure group system number area play game style
player [1237]	game team | year play | number century court field sport
area [1236]	region | part field body number | center form system
part [1177]	role area | surface water number year game field type
work [1171]	play | game study form number music skill group art
skill [1016]	development work variety role method number technique activity
association [955]	group study number role development game work | government
year [917]	century time player | game part play group court | number
group [799]	number form species activity game | association type area
number [793]	variety | game amount group diversity area part form type
skin [774]	surface cell body | material water area tissue form soil wall
fish [766]	animal bird | resource plant life organism body land water
power [745]	state authority number right force control order energy
time [736]	year | number game part record area player | field court
force [713]	army power government | group area number system state
body [658]	area | group part variety skin stream system fish particle
development [638]	study growth problem change role field program part state skill
structure [626]	form system function surface | layer feature pattern area
material [623]	substance particle product rock soil mineral compound energy
law [618]	supreme-court | rule right program legislation control
type [609]	variety form kind number | characteristic part group source
life [606]	activity population group theater art tradition history fish
process [579]	change product number structure heat development game material
sport [577]	game | competition interest | baseball industry player
point [572]	temperature number amount | line part player game star area
industry [567]	product | center activity | source sport development
style [566]	art | artist form tradition work theater architecture music
plant [557]	animal cell production | type variety kind energy number fish
temperature [556]	pressure level | amount flow heat rate | depth density
center [553]	industry area city part development | resource activity
century [551]	year | player work number game | sport music time king
state [548]	country government power | development city area system
de [543]	herzog-herzogin-furst-furstin-prinz des | van du da son
activity [533]	function group interest life industry | event resource
line [531]	track number pattern point type | ball area year form surface
region [518]	area | land | ocean part center sea soil coast water basin
art [515]	architecture artist style | literature tradition music
amount [509]	energy temperature pressure quantity level supply | heat
member [500]	man player school part organization leader association support
level [489]	temperature | pressure amount | flow number characteristic
fight [488]	war sport game year | player country hunt effort government
species [483]	animal | bird group type | population area form variety

SPORTS. Semantic Clusters.

See Page 126

image *as* light	energy
image *as* signal	data
importance *as* knowledge	attention
importance *as* popularity	attention success
increase *as* decrease	difference
increase *as* difference	change
increase *as* variation	change difference
individual *as* student	child
industry *as* activity	center
industry *as* center	development
industry *as* equipment	product
industry *as* interest	activity sport
industry *as* product	source production
infection *as* disorder	disease
influence *as* force	power
influence *as* importance	success
institution *as* office	official
instrument *as* device	method
interest *as* activity	industry
interest *as* diversity	variety
interest *as* industry	sport
interest *as* sport	industry
ion *as* compound	molecule mineral metal
ion *as* hydroxide	chloride sulfate
ion *as* metal	compound mineral
ion *as* mineral	compound metal
ion *as* salt	compound mineral
iron *as* glass	metal
iron *as* metal	mineral ion
iron *as* mineral	metal substance ion
iron *as* steel	metal
issue *as* concern	problem
issue *as* dispute	case
judge *as* authority	justice
judge *as* council	justice governor
judge *as* governor	justice
judge *as* officer	official
judge *as* official	authority
jurisdiction *as* responsibility	reform
justice *as* authority	judge
justice *as* council	judge governor
justice *as* governor	judge
justice *as* judge	authority
kind *as* characteristic	type
kind *as* variety	type plant
king *as* queen	kingdom
kingdom *as* queen	king
km *as* caspian	mi centimeter sec oceans-arctic
km *as* centimeter	mi ft

SPORTS. Semantic Clusters. Two-word Terms.

See Page 145

gold-medal *as* basketball-team	silver-medal college-football
gold-medal *as* bronze-medal	silver-medal
gold-medal *as* football-player	tennis-player
gold-medal *as* petroleum-field	soccer-player
gold-medal *as* science-fiction	world-war-ii
gold-medal *as* silver-medal	bronze-medal basketball-team
gold-medal *as* soccer-player	tennis-player petroleum-field
gold-medal *as* tennis-player	football-player college-football soccer-player
land-surface *as* deg-f	deg-c
land-surface *as* sq-km	sq-mi
land-surface *as* sq-mi	sq-km
land-surface *as* surface-temperature	water-vapor
land-surface *as* water-surface	deg-f deg-c
land-surface *as* water-vapor	deg-c
louis-xiv *as* century-ad	court-ballet
louis-xiv *as* church-music	court-ballet duc-de mid-th-century
louis-xiv *as* court-ballet	century-ad
louis-xiv *as* duc-de	court-ballet verse-play
louis-xiv *as* mid-th-century	court-ballet duc-de church-music
natural-gas *as* building-material	wood-product motor-vehicle forest-product
natural-gas *as* citrus-fruit	sugar-beet tobacco-product
natural-gas *as* forest-product	wood-product transportation-equipment
natural-gas *as* motor-vehicle	sugar-beet wood-product transportation-equipment
natural-gas *as* sugar-beet	tobacco-product citrus-fruit
natural-gas *as* wood-product	sugar-beet transportation-equipment tobacco-product
oil-field *as* oil-refinery	petroleum-industry
oil-field *as* petroleum-industry	oil-refinery steel-mill
saline-water *as* deuterium-oxide	sodium-hydroxide
saline-water *as* surface-water	deg-f
salt-water *as* cooler-water	game-fish
salt-water *as* game-fish	marine-water
salt-water *as* marine-life	shallow-water
salt-water *as* marine-water	shallow-water
salt-water *as* water-supply	surface-water
sea-level *as* ice-sheet	atlantic-coast ice-surface
sea-level *as* land-surface	deg-f deg-c water-vapor
shallow-water *as* coral-reef	marine-water
shallow-water *as* marine-life	salt-water water-table
shallow-water *as* marine-water	salt-water
shallow-water *as* ocean-water	coral-reef
state-court *as* county-court	trial-court district-court state-law criminal-case
state-court *as* court-decision	court-system trial-court
state-court *as* court-system	trial-court district-court court-decision
state-court *as* criminal-case	court-system trial-court district-court
state-court *as* district-court	court-system trial-court criminal-case
state-court *as* state-law	state-legislature district-court criminal-case
state-court *as* state-legislature	court-system state-law court-decision
state-court *as* superior-court	trial-court district-court criminal-case court-decision

SPORTS. First-Pass Thesaurus. *See Page 131.*

area :: [1235 contexts, frequency rank 10] SPORTS *Relat.* form, part; system, field; city, number, body, center, region. *Vbs.* increase, ski, fish, find, divide, cover, use, specialize, inhabit, give, reduce, play. *Exp.* surface area (cf. digestive tract, sq km), land area (cf. surface water, surface feature).

association :: [955 contexts, frequency rank 14] SPORTS *Relat.* work, game; government, development, study, number, group. *Vbs.* find, begin, establish, form, found, continue, organize, join, use, result, lead, know. *Exp.* trade association (cf. party split, state law), loan association (cf. graduate education, sea water).

body :: [658 contexts, frequency rank 23] SPORTS *Relat.* fish, skin; group, system, area; movement, particle, stream, variety. *Vbs.* enter, fall, create, protect, consist, use, support, propel, occupy, lie, govern, cover. *Exp.* body surface (cf. computer field, food supply), water body (cf. ice age, sq mi), body temperature (cf. water loss, water balance), body water (cf. body temperature, sodium ion), body cavity (cf. surface layer, road surface).

center :: [553 contexts, frequency rank 37] SPORTS *Relat.* region, activity, development, industry; role, part, area; port, city. *Vbs.* manufacture, locate, begin, play, lie, lead, fish, diversify, contain.

century :: [552 contexts, frequency rank 38] SPORTS *Relat.* game, player, year; beginning, king. *Vbs.* develop, use, begin, appear, date, flourish, play, build, write, produce, continue, start.

court :: [3114 contexts, frequency rank 4] SPORTS *Relat.* field; water; time, government, year, game, player, play, supreme-court. *Vbs.* hold, rule, declare, use, uphold, establish, appoint, speak, try, serve, decide, call. *Exp.* court painter (cf. court poet, imperial court), court decision (cf. trial court, state court), district court (cf. criminal case, superior court), court system (cf. state court, criminal case), state court (cf. state legislature, court system), imperial court (cf. court painter, court poet), trial court (cf. criminal case, court decision), law court (cf. middle class, ball court), superior court (cf. circuit court, district court), court poet (cf. portrait painter, court painter).

development :: [635 contexts, frequency rank 24] SPORTS *Relat.* part, field, role; center, state, program, change, problem, growth, study. *Vbs.* lead, begin, influence, use, encourage, result, remain, make, know, give, follow, facilitate.

field :: [3646 contexts, frequency rank 2] SPORTS *Relat.* surface; water; development, part, court, play, game, form, area, system. *Vbs.* apply, use, produce, generate, enter, dominate, create, relate, develop, specialize, find, move. *Exp.* oil field (cf. natural gas, petroleum industry), field marshal (cf. greenhouse effect, basketball player), field crop (cf. sugar beet, rice field), petroleum field (cf. motion picture, iron ore), field strength (cf. radio wave, flowering plant), field line (cf. social role, quantum mechanic), field work (cf. amateur athlete, root system), field goal (cf. valuable player, court intrigue), rice field (cf. field crop, aquatic life), field study (cf. cell membrane, game theory).

fish :: [766 contexts, frequency rank 19] SPORTS *Relat.* water; ship, land, body, organism, life, bird, plant, animal. *Vbs.* find, swim, fish, fly, hunt, eat, develop, catch, know, feed, resemble, make. *Exp.* game fish (cf. atlantic coast, salt water), sport fish (cf. marine water, game fish), marine fish (cf. marine water, life form), food fish (cf. air bladder, game species).

force :: [713 contexts, frequency rank 22] SPORTS *Relat.* power; system; influence, state, government, army. *Vbs.* exert, arm, use, defeat, fight, match, join, employ, develop, act, organize, oppose. *Exp.* work force (cf. career education, stroke volume), labor force (cf. land surface, natural resource), air force (cf. fighter pilot, world war ii).

6.18 TIME

Name	:	TIME
Size	:	1.5 Megabyte
Documents	:	425 (Average = 676 words)
Words	:	287 K
Unique words	:	22 K
Source	:	IR testbed (ftp'ed from ftp.cs.cornell)
Description	:	Foreign Affairs articles from TIME (early 60's)
Queries	:	83 (Average = 17 words)

Sample Text:

- the allies after nassau in december 1960, the u.s . first proposed to help nato develop its own nuclear strike force . but europe made no attempt to devise a plan . last week, as they studied the nassau accord between president kennedy and prime minister macmillan, europeans saw emerging the first outlines of the nuclear nato that the u.s . wants and will support . it all sprang from the anglo-u.s . crisis over cancellation of the bug-ridden skybolt missile, and the u.s . offer to supply britain and france with the proved polaris (time, . . .
- the road to jail is paved with nonobjective art since the kremlin's sharpest barbs these days are aimed at modern art and " western espionage, " it was just a matter of time before the kgb's cops would turn up a victim whose wrongdoings combined both evils . he turned out to be a leningrad physics teacher whose taste for abstract painting allegedly led him to join the u.s . spy service . police said they first spotted the teacher, one rudolf friedman, as he muttered uncomplimentary remarks about socialist realism while strolling through leningrad's russian museum . . . .

sample queries :

- u.s . policy toward the new regime in south viet nam which overthrew president diem .
- number of troops the united states has stationed in south viet nam as compared with the number of troops it has stationed in west germany .
- growing controversy in southeast asia over the proposed creation of a federation of malaysia .
- the united states has warned it would limit its united nations payments to the level of its regular assessment if nations now in arrears fail to pay up . what issues are involved in these nations' being in arrears .
- persons involved in the viet nam coup .

TIME (1500K) : SEXTANT results, 50 most frequent words

word [*Contexts*]	*Groups of closest words. (See page 50)*
week [970]	leader government minister man day year | official khrushchev
government [716]	leader | party regime year week man | minister president
minister [659]	week leader officer president government premier official
party [637]	government | force man year army time country | nation
year [545]	government month man week time | party day leader troop
gaulle [508]	khrushchev macmillan nas | diem nikita nhu china thing man
man [503]	leader week government year | party troop communist people
leader [477]	government union | member people man week minister force
force [443]	troop | army leader party war defense nation deterrent fleet
war [373]	force crisis control campaign troop diem struggle people policy
nation [340]	country state ally troop | britain force government republic
time [312]	year month day khrushchev people party country support week
official [294]	diplomat troop authority officer | agent chief khrushchev
people [292]	leader police | government troop man war state time family
nam [288]	africa | cambodia laos turkey malaya europe | struggle cong
khrushchev [275]	gaulle britain peke diem | china kennedy government official
army [269]	force | soldier defense police party war official delegation
day [266]	week month time year man night | home government china life
country [262]	nation | party office troop diplomat time people state leader
troop [253]	force | soldier police nation | official officer war leader
communist [240]	friend government | man official citizen peke troop press
nas [227]	gaulle wilson republic | people diplomat government russian
china [223]	khrushchev ally peke britain gaulle nam russia regime
police [222]	troop people | army minister agent regime government leader
month [221]	year time day | week election term government house country
diem [219]	buddhist kennedy khrushchev macmillan adviser nhu peke general
officer [218]	soldier commander troop chief minister official leader
market [215]	trade europe | nato currency britain page territory
union [214]	europe | unity nationalist federation idea african boss
state [212]	nation | people republic minister city friend socialist party
power [211]	rule | control leader independence weapon force majority war
way [208]	vote time hope run ruler policy side army friend point
germany [207]	russia france german berlin suspect secretary berliner
president [207]	premier | minister government boss strongman deputy regime
policy [203]	defense relation war | claim pact cooperation plan support
britain [195]	khrushchev russia nation china rahman | macmillan gaulle hope
home [192]	house office car day south part france | team hope government
house [192]	home palace time friend car moscow building place | city
regime [192]	government leader | rule kassem president diem scheme china
member [191]	leader | chairman chief session | socialist secretary
cent [184]	parliament majority year population germany drop share source
meeting [181]	conference session | boss talk leader week visit secretary
conference [180]	meeting | talk council session union agreement diem congress
line [174]	relation violation conversation force crowd control khrushchev
plan [173]	program proposal scheme effort | policy federation government
control [170]	war power security defense rule troop system independence law
end [169]	visit deputy stop ambassador buddhist | eye ceremony family
election [167]	assembly | france aim merchant premier | month peke session
viet [167]	position | africa | china african | post east
world [164]	europe floor ally moscow nation china city union front france

TIME. Query Experiments Results

See Page 105

TIME

	base	DOC	SEXT	stem	fam	S+fam	S+f+stem
	PRECISION						
Recall: 10	0.750	0.708	0.753	0.768	0.727	0.747	0.749
Recall: 20	0.742	0.697	0.748	0.758	0.722	0.735	0.734
Recall: 30	0.733	0.686	0.727	0.745	0.713	0.735	0.731
Recall: 40	0.729	0.669	0.721	0.738	0.699	0.719	0.713
Recall: 50	0.705	0.647	0.698	0.718	0.673	0.689	0.684
Recall: 60	0.611	0.566	0.608	0.626	0.606	0.606	0.605
Recall: 70	0.585	0.551	0.582	0.602	0.588	0.590	0.595
Recall: 80	0.574	0.530	0.569	0.589	0.575	0.582	0.585
Recall: 90	0.534	0.470	0.528	0.547	0.535	0.541	0.548
Average	0.663	0.614	0.659	0.677	0.649	0.660	0.661
Better	---	28	8	23	27	25	30
Same	---	20	56	42	35	29	27
Worse	---	35	19	18	21	29	26
	RECALL						
At 5 docs:	0.35	0.33	0.34	0.35	0.36	0.35	0.36
At 10 docs:	0.27	0.27	0.27	0.28	0.27	0.27	0.28
At 15 docs:	0.21	0.21	0.21	0.21	0.21	0.21	0.22
At 20 docs:	0.17	0.17	0.17	0.17	0.17	0.17	0.17
At 25 docs:	0.14	0.14	0.14	0.14	0.14	0.14	0.14
Better at 15	---	7	2	3	6	7	9
Same at 15	---	70	78	78	74	73	72
Worse at 15	---	6	3	2	3	3	2

TIME --- BEST IMPROVEMENTS (*see page 105*)

Base Query	*Augmented Query*	*change*
king sign power state free rein half-brother feisal reform rule	king sign power state free rein half-brother feisal **saud swing saudi** reform **science structure reform-minded scientific** rule	0.333 to 1.000
alternative offer u force withdraw congo	alternative **determination** offer force **troop** withdraw congo **congolese**	0.194 to 0.704
effort three-nation international control commission indo-china try stop fight flare laos	effort three-nation international control commission indo-china try stop fight flare laos **lao laotian**	0.489 to 0.833
indian fear communist chinese invasion	indian **india** fear communist chinese **china** invasion	0.440 to 0.773
agreement syria iraq full economic unity close economic cooperation	agreement syria **syrian** iraq **iraqi** full economic **economy** unity close economic **economy** cooperation **relation**	0.733 to 1.000

TIME --- WORST RESULTS

Base Query	*Augmented Query*	*change*
team survey public opinion north borneo sarawak question join federation malaysia	team survey **poll** public opinion north borneo sarawak question join federation **coalition** malaysia **miracle malay malaya malayan**	1.000 to 0.574
conflict israel arab neighbor	conflict israel **israeli** arab **republic** neighbor	1.000 to 0.500
withdrawal sultanate brunei propose federation malaysia	withdrawal sultanate brunei propose federation **coalition** malaysia **miracle malay malaya malayan**	1.000 to 0.426
talk hold east germany premier khrushchev leader east european satellite country	talk hold east germany **german** premier khrushchev leader **government** east european **europe** satellite **hierarchy** country	1.000 to 0.250

TIME. Semantic Clusters.

See Page 126

khrushchev *as* britain	china
khrushchev *as* china	gaulle
khrushchev *as* diem	gaulle
khrushchev *as* kennedy	diem
khrushchev *as* official	week
khrushchev *as* peking	diem china
leader *as* government	week
leader *as* man	government week
leader *as* minister	government week
leader *as* people	government man
leader *as* premier	minister
leader *as* regime	government
life *as* power	war
macmillan *as* diem	gaulle
macmillan *as* king	peking
macmillan *as* peking	diem king
macmillan *as* wilson	diem
man *as* communist	government
man *as* day	week year
man *as* government	week
man *as* leader	government week
man *as* party	government
man *as* people	leader government
man *as* year	government week party
meeting *as* session	conference
meeting *as* talk	conference
member *as* chairman	boss
member *as* leader	minister
minister *as* government	week
minister *as* leader	week government
minister *as* officer	official
minister *as* official	week
minister *as* premier	leader president
minister *as* president	government
month *as* day	year time week
month *as* house	time
month *as* time	year
month *as* year	week government
nam *as* cambodia	laos
nation *as* people	government
nation *as* state	people
nation *as* troop	force people
office *as* home	house
officer *as* chief	official
officer *as* commander	soldier chief
officer *as* official	minister
officer *as* soldier	troop
officer *as* troop	official
official *as* agent	diplomat

TIME. Semantic Clusters. Two-word Terms.

See Page 145

iron-curtain *as* east-berlin	west-berlin east-germany
iron-curtain *as* east-german	east-germany west-german
iron-curtain *as* east-germany	west-berlin west-german
iron-curtain *as* west-berlin	east-germany air-force
king-hassan *as* arab-union	north-africa state-department arab-world
king-hassan *as* arab-world	middle-east
king-hassan *as* north-africa	arab-union state-department arab-world
king-hussein *as* arab-union	king-saud arab-world baath-party michel-aflak
king-hussein *as* arab-unity	arab-world
king-hussein *as* arab-world	saudi-arabia arab-unity
king-hussein *as* king-paul	state-department
king-hussein *as* king-saud	arab-union
king-hussein *as* michel-aflak	arab-unity baath-party
king-paul *as* king-hussein	state-department
king-paul *as* market-membership	queen-frederika
king-paul *as* opinion-poll	market-membership labor-government
king-paul *as* queen-frederika	state-visit market-membership
king-paul *as* state-department	king-hussein
kong-le *as* mao-tse-tung	pathet-lao
kong-le *as* pathet-lao	viet-nam viet-cong president-kennedy
kong-le *as* viet-cong	viet-nam west-germany
labor-party *as* general-election	harold-wilson harold-macmillan
labor-party *as* harold-macmillan	ne-win south-africa
labor-party *as* harold-wilson	harold-macmillan ne-win
labor-party *as* hugh-gaitskell	harold-wilson general-election
labor-party *as* profumo-case	harold-wilson
labor-party *as* south-africa	viet-nam
lei-feng *as* past-year	hong-kong
mao-tse-tung *as* communist-party	nikita-khrushchev
mao-tse-tung *as* kong-le	pathet-lao
mao-tse-tung *as* pathet-lao	communist-party kong-le
mao-tse-tung *as* sino-soviet-split	test-ban
mao-tse-tung *as* soviet-union	communist-party nikita-khrushchev
mao-tse-tung *as* test-ban	soviet-union
michel-aflak *as* arab-union	baath-party arab-world king-hussein
michel-aflak *as* arab-unity	arab-world
michel-aflak *as* baath-party	arab-unity saudi-arabia communist-party
michel-aflak *as* king-hussein	arab-unity saudi-arabia arab-world
middle-east *as* abdul-rahman	viet-nam
middle-east *as* arab-unity	arab-world
middle-east *as* arab-world	saudi-arabia arab-unity
middle-east *as* baath-party	arab-world saudi-arabia arab-unity
middle-east *as* saudi-arabia	west-germany
middle-east *as* southeast-asia	viet-nam abdul-rahman
middle-east *as* world-war	abdul-rahman west-germany
miss-x *as* east-berlin	state-visit iron-curtain
ne-win *as* harold-macmillan	labor-party abdul-rahman
ne-win *as* harold-wilson	labor-party harold-macmillan
nikita-khrushchev *as* communist-party	viet-nam

TIME. First-Pass Thesaurus. *See Page 131.*

army :: [269 contexts, frequency rank 17] TIME *Relat.* government, party, force; union, unity, police, defense, soldier. *Vbs.* own.

britain :: [195 contexts, frequency rank 35] TIME *Relat.* china; nation, gaulle, khrushchev; hope, macmillan, rahman, russia. *Vbs.* say, veto, own, insist, commit, call. *Fam.* britannia, british.

china :: [223 contexts, frequency rank 23] TIME *Relat.* diem; nam, gaulle, khrushchev; organization, regime, russia, britain, ally, peking. *Vbs.* say, announce. *Fam.* chinese.

communist :: [240 contexts, frequency rank 21] TIME *Relat.* man, government; friend. *Vbs.* break. *Exp.* communist party (cf. soviet union, president kennedy), communist china (cf. mao tse tung, nikita khrushchev).

country :: [262 contexts, frequency rank 19] TIME *Relat.* time; party, nation; strike, office. *Vbs.* leave, say, run, own, continue, order, hold, divide.

day :: [266 contexts, frequency rank 18] TIME *Relat.* time; man, government, year, week; home, night, month. *Vbs.* spend, arrive, recall. *Fam.* day-long.

diem :: [219 contexts, frequency rank 26] TIME *Relat.* gaulle, khrushchev; wilson, nhu, general, peking, adviser, macmillan, kennedy, buddhist. *Vbs.* show. *Exp.* diem government (cf. quang duc, buddhist monk), president diem (cf. quang duc, guerrilla war).

force :: [443 contexts, frequency rank 9] TIME *Relat.* leader; party; problem, fleet, defense, army, nation, war, troop. *Vbs.* arm, own, join, create, use. *Exp.* air force (cf. southeast asia, viet cong), police force (cf. security council, east germany), strike force (cf. polaris submarine, north africa).

gaulle :: [508 contexts, frequency rank 6] TIME *Relat.* man; nenni, thing, china, nhu, nikita, nasser, diem, macmillan, khrushchev. *Vbs.* make, say, kill, want, turn, try, know.

germany :: [207 contexts, frequency rank 33] TIME *Relat.* suspect, berliners, berlin, german, europe, secretary, france, russia. *Vbs.* look. *Exp.* west germany (cf. de gaulle, west german), east germany (cf. east berlin, east german). *Fam.* german.

government :: [716 contexts, frequency rank 2] TIME *Relat.* minister; week; communist, people, year, president, regime, party, man, leader. *Vbs.* overthrow, say, promise, give, resign, recognize, ask, want, try, topple, take, support. *Exp.* diem government (cf. quang duc, buddhist monk), government official (cf. von horn, kennedy administration), labor government (cf. opinion poll, way back), government office (cf. king hussein, socialist party).

home :: [192 contexts, frequency rank 36] TIME *Relat.* house; day; team, part, south, car, office. *Vbs.* return, take, send, remain.

house :: [192 contexts, frequency rank 36] TIME *Relat.* month, home; time; place, moscow, car, palace. *Vbs.* live, lie.

khrushchev :: [275 contexts, frequency rank 16] TIME *Relat.* time, official; week, gaulle; kennedy, diem, peking, china, britain. *Vbs.* say, come, turn, warn, want, suggest, declare.

leader :: [477 contexts, frequency rank 8] TIME *Relat.* force, man; minister, week, government; premier, regime, people, member. *Vbs.* say, warn, own.

6.19 XRAY

Name	:	XRAY
Size	:	5.88 megabyte
Documents	:	5804 (Average = 150 words)
Words	:	880 K
Unique words	:	7530 (!!)
Description	:	Hospital X-ray data

Sample Text:

- There is further resolution of the right pleural effusion since 12-12-91. Minimal residual pleural thickening is evident. Parenchymal opacity in the right middle lobe has also diminished. The left lung is clear.

  Apperance of the chest is approaching a new baseline for this patient following resolution of right pleural effusion.
- The heart is not enlarged. The lung fields are clear of infiltrates or cavitary disease. The visualized portion of thoracic cage is intact.

  Heart and lungs within normal limits. No evidence of tuberculosis.

  There is no evidence of pneumonia. Tortuous aorta noted. There are no previous films available for comparison. No evidence of infiltrate.
- A PA view only was obtained due to the patient's pregnancy. Given this, the cardiovascular silhouette is normal. The lungs are clear.

  The heart is normal in size. No pulmonary parenchymal or pleural abnormality is noted.

  No evidence of pulmonary infiltrate.

  The lungs are clear. Heart size is normal. No bony abnormality is seen.

  Normal chest.
- PA and lateral views are compared to the previous exam dated 4-26-91 and 3-10-91. The lungs are clear without evidence of congestive heart failure or pneumonia. The cardiomediastinal contours are within normal limits. There has been interval resolution of the previously noted right lower lobe and left basilar infiltrates. There is no evidence of pleural effusion.

  No evidence of pneumonia, resolution of previously noted right lower lobe and left lower lobe infiltrates.
- PA view only is interpreted without comparison films. The cardiomediastinal silhouette is within normal limits. The lungs are clear bilaterally and the pleural surfaces appear normal.

XRAY (K) : SEXTANT results, 50 most frequent words

word [*Contexts*]	*Groups of closest words. (See page 50)*
view [19121]	film | examination | study exam change opacity lung
effusion [13555]	pneumothorax | edema | opacity | pneumonia change fluid
lung [12740]	effusion | appearance edema silhouette | thickening
film [11338]	view examination | study | exam | radiograph change
change [8926]	disease | effusion abnormality | atelectasis opacity
opacity [8256]	density | atelectasis | effusion | mass consolidation
tube [7950]	catheter | effusion | density atelectasis level clip
silhouette [7657]	heart | appearance | contour cardiomegaly | mediastinum
line [7246]	effusion | clip opacity density mass atelectasis lung
atelectasis [6562]	consolidation | opacity edema | density pneumothorax
study [6416]	exam examination | film | image chest radiograph | x_ray
disease [6166]	abnormality change | consolidation | effusion opacity edema
lobe [6086]	base | effusion | zone thickening hemithorax | disease
chest [5900]	study | silhouette projection position | mediastinum
evidence [5732]	compatible change | effusion | edema lung abnormality lobe
abnormality [5061]	disease | density change consolidation | mass effusion
pneumothorax [4501]	effusion | fluid atelectasis | abnormality consolidation
catheter [4318]	tube | clip pneumothorax wire | density thickening
edema [4279]	effusion failure atelectasis | pneumonia consolidation
position [4001]	end | chest appearance region | present volume placement
left [3871]	consolidation | edema area disease thickening | compatible
base [3709]	lobe | volume area zone appearance | hemithorax
right [3674]	edema | abnormality disease | compatible failure | scar
failure [3595]	edema | chf pneumonia | consolidation cardiomegaly
tip [3254]	region | level | placement clip present portion
examination [2992]	study | exam | film | radiograph | x_ray projection
heart [2730]	silhouette | cardiomegaly | mediastinum shadow | volume
volume [2650]	base edema improvement | consolidation atelectasis effusion
size [2634]	cardiomegaly | volume appearance | enlargement limit
exam [2596]	study | examination | radiograph x_ray | film show image
appearance [2594]	base silhouette size change abnormality | pattern opacity
density [2473]	opacity | mass consolidation | thickening area atelectasis
configuration [2338]	contour | enlargement | limit size | silhouette border
fracture [2329]	abnormality | pneumothorax change effusion disease nodule
pneumonia [2251]	consolidation edema | effusion failure | scar | disease
angle [2215]	sulcus | base hemidiaphragm field | elevation | hilum
limit [2163]	size | configuration | difference top | cardiomegaly
thickening [2119]	scar | density | consolidation calcification pneumothorax
clip [2004]	staple | wire mass emphysema | calcification drain density
redistribution [1968]	cephalization cardiomegaly | haziness | engorgement
aorta [1894]	enlargement silhouette size cardiomegaly colon heart
comparison [1865]	examination | evaluation scan | day pa radiograph x_ray
consolidation [1840]	atelectasis | opacification pneumonia | density edema
sulcus [1797]	angle | hemidiaphragm border | base zone hemithorax field
enlargement [1737]	cardiomegaly | prominence | failure redistribution
mass [1730]	nodule | opacity abnormality | adenopathy | area
region [1706]	mass area | density portion base | abnormality aspect
portable [1683]	image | pa | scan ct series position inspiration report
air [1622]	gas | thickening density abnormality | opacification
tissue [1601]	emphysema | clip collection | rib heart adenopathy area

XRAY. Semantic Clusters.

See Page 126

ileus *as* obstruction	bowel
image *as* ap	portable examination study chest
image *as* exam	examination study chest
image *as* examination	study chest
image *as* kub	portable ap radiograph x-ray
image *as* portable	chest exam
image *as* radiograph	portable examination study chest exam
image *as* x-ray	radiograph examination study exam
improvement *as* cardiomegaly	redistribution
improvement *as* clearing	decrease resolution increase progression development
improvement *as* decrease	increase
improvement *as* development	decrease resolution degree
improvement *as* engorgement	cardiomegaly redistribution
improvement *as* progression	clearing development
improvement *as* resolution	decrease increase
increase *as* clearing	decrease improvement resolution
increase *as* decrease	improvement
increase *as* improvement	decrease resolution cardiomegaly
increase *as* prominence	cardiomegaly redistribution
increase *as* resolution	decrease improvement
infusion *as* hickman	ij port-a-cath
infusion *as* ij	port-a-cath
infusion *as* quinton	hickman approach ij port-a-cath transvenous
infusion *as* swan-ganz-catheter	approach
infusion *as* transvenous	quinton hickman
interval *as* past	yesterday
interval *as* progression	improvement
interval *as* yesterday	day
joint *as* cartilage	space osteoarthritis knee
knee *as* ankle	hip
knee *as* foot	hip shoulder ankle decubitus
knee *as* hip	prosthesis
knee *as* prostheses	prosthesis replacement
knee *as* shoulder	hip
left *as* change	effusion lung
left *as* lobe	lung
left *as* pneumothorax	effusion atelectasis abnormality
left *as* right	effusion lung change lobe opacity
lesion *as* adenopathy	mass nodule pneumothorax
lesion *as* area	density
lesion *as* mass	abnormality density
lesion *as* masse	metastasis
lesion *as* metastasis	adenopathy
lesion *as* nodule	mass abnormality density pneumothorax
level *as* fluid	air
limit *as* configuration	size
limit *as* size	silhouette
line *as* catheter	tube
line *as* end	catheter tube tip

XRAY. Semantic Clusters. Two-word Terms.

See Page 145

ij-line *as* catheter-tip	line-tip interval-removal
ij-line *as* ge-junction	tracheostomy-tube good-position
ij-line *as* good-position	interval-removal
ij-line *as* hickman-line	line-tip catheter-tip interval-placement
ij-line *as* interval-placement	tracheostomy-tube line-tip catheter-tip
ij-line *as* line-tip	good-position interval-removal
ij-line *as* pa-line	line-tip catheter-tip ge-junction
ij-line *as* tracheostomy-tube	line-tip good-position interval-removal
infrahilar-region *as* lobe-opacification	retrocardiac-area
infrahilar-region *as* lung-region	retrocardiac-area
infrahilar-region *as* retrocardiac-area	retrocardiac-region lung-region air-bronchogram
infrahilar-region *as* suture-line	cavitary-lesion
interval-change *as* air-space	lung-base lung-zone
interval-change *as* chest-film	air-space lung-base cardiomediastinal-silhouette
interval-change *as* chest-wall	lung-base chest-film lung-zone chest-tube
interval-change *as* lobe-opacity	air-space lung-base lung-zone lobe-pneumonia
interval-change *as* lobe-pneumonia	air-space lung-zone
interval-change *as* lung-base	chest-tube cardiomediastinal-silhouette
interval-change *as* lung-zone	lung-base chest-tube
interval-change *as* status-post	chest-tube
interval-decrease *as* interval-development	interval-improvement interval-resolution
interval-decrease *as* interval-improvement	lobe-consolidation lung-volume lobe-opacity
interval-decrease *as* interval-increase	interval-resolution
interval-decrease *as* interval-resolution	interval-improvement lobe-consolidation
interval-decrease *as* lobe-infiltrate	interval-resolution lobe-consolidation
interval-development *as* interval-increase	interval-resolution interval-decrease
interval-development *as* interval-resolution	interval-improvement lobe-pneumonia
interval-development *as* lobe-collapse	lobe-consolidation
interval-development *as* lobe-opacity	lobe-pneumonia air-space
interval-development *as* lobe-pneumonia	air-space
interval-development *as* retrocardiac-opacity	interval-resolution lobe-consolidation
interval-improvement *as* air-space	lung-zone
interval-improvement *as* interval-resolution	interval-decrease lobe-pneumonia
interval-improvement *as* lobe-consolidation	lobe-pneumonia lobe-opacity air-space
interval-improvement *as* lobe-opacity	lobe-pneumonia air-space lung-zone
interval-improvement *as* lobe-pneumonia	air-space lung-zone
interval-improvement *as* lung-volume	lobe-pneumonia lobe-opacity
interval-increase *as* interval-development	interval-resolution interval-decrease
interval-increase *as* interval-resolution	interval-decrease interval-improvement
interval-placement *as* catheter-tip	line-tip interval-removal
interval-placement *as* cavoatrial-junction	line-tip good-position
interval-placement *as* good-position	interval-removal ng-tube
interval-placement *as* hickman-line	line-tip catheter-tip ij-line cavoatrial-junction
interval-placement *as* ij-line	line-tip interval-removal good-position
interval-placement *as* interval-removal	ng-tube
interval-placement *as* line-tip	interval-removal good-position
interval-placement *as* side-vent	good-position tracheostomy-tube

XRAY. First-Pass Thesaurus. *See Page 131.*

abnormality :: [5061 contexts, frequency rank 16] BW *Relat.* disease; opacity, effusion, change; pneumothorax, thickening, nodule, consolidation, mass, density. *Vbs.* note, show, identify, demonstrate, follow, appear, reveal, side, infiltrate, unchange, detect, describe. *Exp.* contour abnormality (cf. pa projection, interval resolution), rib abnormality (cf. rib lesion, callus formation), lobe abnormality (cf. lobe granuloma, lobe density), lung abnormality (cf. lung infiltrate, tumor recurrence).

angle :: [2215 contexts, frequency rank 36] BW *Relat.* border, mediastinum, field, elevation, sulcus. *Vbs.* blunt, note, remain, exclude, suggest, cut, unchange, obscure, increase, resolve, appear. *Exp.* cp angle (cf. retrocardiac density, pa projection), angle region (cf. retrocardiac space, cp angle).

appearance :: [2594 contexts, frequency rank 31] BW *Relat.* silhouette, size; finding, pattern. *Vbs.* unchange, note, follow, improve, suggest, compare, change, remain, transplant, enlarge, contribute, give.

atelectasis :: [6562 contexts, frequency rank 10] BW *Relat.* lobe; opacity, effusion; pneumothorax, scar, thickening, density, edema, pneumonia, consolidation. *Vbs.* note, represent, infiltrate, unchange, increase, improve, continue, leave, associate, compare, persist, suggest.

base :: [3709 contexts, frequency rank 22] BW *Relat.* atelectasis, lung, lobe; volume, apex, area, zone, hemithorax, scar. *Vbs.* note, infiltrate, elevate, remain, exclude, leave, increase, unchange, improve, scar, represent, persist.

catheter :: [4318 contexts, frequency rank 18] BW *Relat.* tube, line; stent, placement, end, tip. *Vbs.* unchange, remove, remain, side, note, reach, place, terminate, show, demonstrate, project, position. *Exp.* catheter tip (cf. line tip, ij line), catheter placement (cf. lung transplantation, line placement).

change :: [8926 contexts, frequency rank 5] BW *Relat.* opacity; lung, effusion; right, left, abnormality, disease. *Vbs.* note, compare, show, reveal, demonstrate, make, represent, occur, unchange, mark, date, describe. *Exp.* interval change (cf. air space, lung base), radiation change (cf. retrocardiac space, lung contusion), lung change (cf. gasseous distention, edema pattern), day change (cf. lung transplantation, comparison film).

chest :: [5900 contexts, frequency rank 14] BW *Relat.* study; view, film; ap, portable, exam, examination, radiograph. *Vbs.* compare, show, unchange, note, overlie, demonstrate, interpret, side, review, reveal, submit, leave. *Exp.* chest tube (cf. ng tube, lung base), chest film (cf. interval change, air space), chest exam (cf. chest examination, chest pa), chest examination (cf. chest pa, lobe density), chest radiograph (cf. lobe nodule, pa film), chest wall (cf. lung zone, interval change), chest view (cf. staple line, lung nodule), chest fluoroscopy (cf. lung nodule, repeat film), chest disease (cf. cardiomediastinal appearance, chest pa), chest pa (cf. chest examination, chest disease).

clip :: [2004 contexts, frequency rank 39] BW *Relat.* thickening; density; contour, surgery, calcification, staple, wire, suture. *Vbs.* note, unchange, project, make, overlie, remove, demonstrate, remain, show, scatter, associate, widen.

configuration :: [2338 contexts, frequency rank 33] BW *Relat.* limit; failure, heart, silhouette, size; border, enlargement, contour. *Vbs.* show, unchange, enlarge, suggest, change.

density :: [2473 contexts, frequency rank 32] BW *Relat.* thickening; atelectasis, abnormality, opacity; calcification, area, opacification, consolidation, nodule, mass. *Vbs.* increase, note, represent, project, unchange, overlie, round, remain, compare, suggest, define, show. *Exp.* bone density (cf. fracture deformity, chest examination), retrocardiac density (cf. air bronchogram, retrocardiac opacity), ossific density (cf. fracture deformity, cartilage calcification), lobe density (cf. chest examination, lobe effusion).

6.20 THESIS

Name	:	THESIS
Size	:	237 kilobyte
Documents	:	77 sections (Average = 570 words)
Words	:	44,000
Unique words	:	4438
Source	:	Chapters 1 to 5 of this book
Description	:	book on semantic discovery

Sample Text:

- Some immediately evident problems of language variability are addressed by any computer system that ignores upper and lower case differences, or that allows truncation of suffixes and prefixes. Such character string manipulations are well-understood and ubiquitously implemented, but only scratch at the surface of the problems natural languages cause.
- Considering these clues as a word's attributes, similarity measures between words can be calculated. Many similarity measures have been defined and used over the past seventy years (Romesburg 1990). The measures take into account the number of attributes that two objects do or do not share, as well as the importance of these attributes for each word.
- Choueka argues that any manually constructed list of two-, three- or four-word terms will not be able to cover the new expressions formed daily in newsprint. He proposes an automatic means of deriving interesting expressions, using the frequency of appearance of the expressions in a large corpus. He proceeds by storing lists of potential expressions appearing more than times in the corpus.
- Since this dictionary was constructed as a learning dictionary, many definitions are of a predictable form, e.g. ''(word): a (word2) that ..." as in the definition given below. {quote} { anaesthetist: a doctor who gives an anaesthetic to a patient } {quote} This definition shows that an {anaesthetist} is type of {doctor.}
- At this point another simple grammar uses the contextual information of English capitalization to join together sequences of words beginning with an uppercase letter, not appearing after a punctuation mark, as a rapid name recognizer.

THESIS (237K) : SEXTANT results, 50 most frequent words

word [*Contexts*]	*Groups of closest words. (See page 50)*
word [773]	relation term context number | pair technique information
technique [262]	method approach | system sextant | measure information
context [241]	attribute | number information relation list | word
corpus [208]	text | data-base document number measure attribute window
phrase [194]	structure relation context unit number | term attribute way
number [186]	context pair | list word frequency information | phrase
relation [179]	context attribute word | pair information similarity phrase
result [178]	sextant | measure list experiment context | technique
list [163]	context number | result pair attribute thesaurus produce
sextant [157]	result technique | attribute patient context system |
information [147]	context knowledge | relation number similarity technique
text [136]	corpus sentence query | data dictionary relation information
noun [133]	verb adjective | attribute | term similarity | relation
system [133]	technique | method approach sextant | research source
term [125]	document word | noun | phrase attribute list query
measure [121]	calculation attribute | result query technique method animal
approach [119]	technique | application information | system method tool
similarity [115]	attribute | context relation noun information query | mean
thesaurus [114]	dictionary structure list | source relation document
pair [101]	number relation | frequency context list word sense document
problem [98]	aspect | drawback text description noun complexity number
sense [97]	synonym pair relation | attribute category similarity
structure [86]	source thesaurus | attribute phrase | list similarity
method [81]	technique | measure system experiment analysis relation test
analysis [76]	experiment extraction discovery | comparison cooccurrence
document [76]	term string | query unit | pair vector thesaurus element
attribute [73]	context | similarity measure relation noun verb | usage
query [71]	way document string measure idea attribute text | similarity
time [65]	percentage relation number | algorithm measure subject
experiment [61]	comparison | analysis module manner application | result
knowledge [61]	information semantics | representation | marker vocabulary
axis [59]	reduction top | group verb cluster attribute relationship
dictionary [59]	thesaurus | source text lexicon information human vocabulary
frequency [58]	window cooccurrence | pair number percentage |
language [57]	natural-language-processing front-ends example concept text
example [55]	algorithm doctor recognition order language part response name
synonym [53]	image | sense | input patient adjective idea measure
adjective [52]	noun | note attribute head appearance | response fall
data [49]	attribute entry text | type window corpus context method
mean [48]	similarity expansion pair weighting context recognizer
unit [48]	pattern | string tag window phrase document sample windowing
source [47]	structure | lexicon thesaurus information | dictionary
level [46]	measure precision | layer improvement weight frequency space
part [46]	vehicle | feature | animal measure | example overlap
group [45]	np hierarchy | variant set axis context subheading distance
retrieval [45]	functionality performance | precision | application
cooccurrence [44]	closeness | presence schutze92 appearance | stem
marker [44]	knowledge constraint structure | process link model
verb [44]	noun | attribute | form response | axis tag manner
application [42]	resource experiment approach extraction power | evaluation

THESIS. Semantic Clusters.

See Page 126

idea *as a* average	meaning
idea *as a* meaning	average
improvement *as a* aspect	limitation change
improvement *as a* limitation	aspect change
information *as a* number	context
information *as a* relation	context technique
information *as a* similarity	context relation
information *as a* source	thesaurus
information *as a* thesaurus	relation
institution *as a* choueka	harvard
institution *as a* iobj	ruge
institution *as a* ruge	iobj
interface *as a* front-ends	power expression
interface *as a* power	front-ends application expression evaluation
interface *as a* resource	application
item *as a* cancer	cat
item *as a* cat	cancer
item *as a* choice	indexer classification
item *as a* classification	choice
item *as a* entity	hierarchy
item *as a* indexer	choice
item *as a* methodology	hierarchy
judgment *as a* effort	choice
judgment *as a* judgement	comparison extraction
judgment *as a* pairing	choice
knowledge *as a* property	extraction
language *as a* text	corpus
list *as a* number	context
list *as a* pair	context number
list *as a* structure	attribute thesaurus
manipulation *as a* count	analysis
manipulation *as a* creation	system
manner *as a* candidate	scheme
manner *as a* order	variety
manner *as a* verb	attribute
marker *as a* constraint	process
marker *as a* primitive	model formula
marker *as a* process	constraint
marker *as a* structure	attribute
match *as a* place	head produce
matrix *as a* space	vocabulary
matrix *as a* vocabulary	space
mean *as a* pair	context relation
mean *as a* similarity	context relation
mean *as a* weighting	scheme
meaning *as a* average	variant idea entry
meaning *as a* idea	average
meaning *as a* spellings	usage

THESIS. Semantic Clusters. Two-word Terms.

See Page 145

individual-word *as* similarity-list	similarity-measure test-bed
individual-word *as* similarity-measure	noun-phrase
individual-word *as* test-bed	language-variability
individual-word *as* word-sense	noun-phrase
information-retrieval *as* average-precision	query-expansion test-bed query-term
information-retrieval *as* language-variability	natural-language
information-retrieval *as* machine-translation	language-variability natural-language
information-retrieval *as* natural-language	noun-phrase
information-retrieval *as* query-expansion	test-bed query-term
information-retrieval *as* query-term	query-expansion language-variability
information-retrieval *as* similarity-measure	noun-phrase
information-retrieval *as* test-bed	language-variability
internal-structure *as* deese-antonym	verb-phrase
internal-structure *as* word-variability	knowledge-poor-method
jaccard-measure *as* judge-similarity	characteristic-vocabulary
jaccard-measure *as* similarity-calculation	similarity-measure
jaccard-measure *as* similarity-measure	information-retrieval
judge-similarity *as* characteristic-vocabulary	query-term
judge-similarity *as* context-point	context-word
judge-similarity *as* context-word	vice-versa context-point
judge-similarity *as* low-frequency-word	frequency-group
judge-similarity *as* vice-versa	context-word context-point
knowledge-poor-approach *as* internal-structure	deese-antonym
knowledge-poor-approach *as* language-understanding	machine-translation language-variability
knowledge-poor-approach *as* lexico-syntactic-pattern	knowledge-structure
knowledge-poor-approach *as* machine-translation	language-understanding language-variability
knowledge-poor-approach *as* world-knowledge	machine-translation language-understanding
knowledge-poor-method *as* knowledge-poor-technique	natural-language
knowledge-poor-method *as* lexico-syntactic-pattern	text-collection knowledge-structure
knowledge-poor-method *as* text-collection	word-variability lexico-syntactic-pattern
knowledge-poor-method *as* word-variability	text-collection
knowledge-poor-technique *as* extraction-technique	knowledge-structure gold-standard
knowledge-poor-technique *as* gold-standard	natural-language extraction-technique
knowledge-poor-technique *as* knowledge-structure	natural-language extraction-technique
knowledge-structure *as* evaluation-technique	extraction-technique characteristic-vocabulary
knowledge-structure *as* extraction-technique	knowledge-poor-technique gold-standard
knowledge-structure *as* gold-standard	knowledge-poor-technique
knowledge-structure *as* knowledge-poor-technique	natural-language
knowledge-structure *as* lexico-syntactic-pattern	thesaurus-enrichment
knowledge-structure *as* thesaurus-enrichment	extraction-technique
language-understanding *as* knowledge-poor-approach	machine-translation world-knowledge
language-understanding *as* language-variability	information-retrieval natural-language
language-understanding *as* machine-translation	language-variability
language-understanding *as* world-knowledge	machine-translation language-variability
language-variability *as* language-understanding	machine-translation
language-variability *as* machine-translation	information-retrieval natural-language
language-variability *as* natural-language	information-retrieval
language-variability *as* query-term	information-retrieval test-bed
language-variability *as* test-bed	information-retrieval individual-word

THESIS. First-Pass Thesaurus. *See Page 131.*

adjective :: [52 contexts, frequency rank 34] THESIS *Relat.* document, attribute, noun; introduction, animal, fall, response, note. *Vbs.* modify, appear, occur, associate, use, consider.

analysis :: [76 contexts, frequency rank 24] THESIS *Relat.* method; measure; count, parser, discovery, extraction, experiment. *Vbs.* perform, use.

approach :: [119 contexts, frequency rank 16] THESIS *Relat.* measure, system; sextant, technique; tool, method, application. *Vbs.* take, use, appear, suggest, process, interest, base.

attribute :: [73 contexts, frequency rank 25] THESIS *Relat.* structure; measure, noun, relation, similarity, context; adjective, association, usage, verb. *Vbs.* share, possess, compare, give, consider.

axis :: [59 contexts, frequency rank 29] THESIS *Relat.* attribute; subheading, verb, group, top, space, reduction. *Vbs.* define, find, create.

context :: [241 contexts, frequency rank 3] THESIS *Relat.* word; pattern, pair, list, information, phrase, number, similarity, relation, attribute. *Vbs.* use, possess, extract, add, compare, examine, share, provide, give, double, derive, appear. *Exp.* context point (cf. vice versa, deese antonym), context pair (cf. similarity judgement, similarity comparison), word context (cf. similarity comparison, sample text), context word (cf. windowing technique, vice versa).

corpus :: [208 contexts, frequency rank 4] THESIS *Relat.* term, document, text. *Vbs.* appear, give, use, extract, hit, treat, test, occur, generate, describe, apply, show.

data :: [49 contexts, frequency rank 35] THESIS *Relat.* corpus, text; definition, entry. *Vbs.* provide, correspond.

dictionary :: [59 contexts, frequency rank 29] THESIS *Relat.* corpus, text, information, thesaurus; encyclopedia, human, lexicon, source. *Vbs.* use, find, create. *Exp.* dictionary entry (cf. word sense, windowing technique), dictionary sense (cf. dictionary definition, language understanding), dictionary definition (cf. similarity pair, head word).

document :: [76 contexts, frequency rank 24] THESIS *Relat.* query; thesaurus, pair, corpus, term; adjective, element, unit, string. *Vbs.* appear, use, rank, index, find, consider.

example :: [55 contexts, frequency rank 32] THESIS *Relat.* language; name, part, doctor, algorithm. *Vbs.* give, show, use, find.

experiment :: [61 contexts, frequency rank 28] THESIS *Relat.* result, method, analysis; scheme, application, manner, module, comparison. *Vbs.* run, perform, describe, use, show, present.

frequency :: [58 contexts, frequency rank 30] THESIS *Relat.* phrase, number, pair; entropy, distribution, cooccurrence, window. *Vbs.* use, appear, accord, rank. *Fam.* frequent

information :: [147 contexts, frequency rank 11] THESIS *Relat.* number, technique, relation, context; thesaurus, source, similarity, knowledge. *Vbs.* extract, use, structure, provide, wad, reap, derive, contain, calculate. *Exp.* information retrieval (cf. query expansion, language variability), extract information (cf. domain knowledge, knowledge structure).

knowledge :: [61 contexts, frequency rank 28] THESIS *Relat.* information; existence, extraction, property, vocabulary, marker, representation, semantics. *Vbs.* structure, generate. *Exp.* knowledge structure (cf. knowledge poor technique, extraction technique), world knowledge (cf. language variability, knowledge poor approach), domain knowledge (cf. extract information, extraction technique). *Fam.* knowledge-poor.

language :: [57 contexts, frequency rank 31] THESIS *Relat.* example; corpus, text; tool, concept, front-ends, natural-language-processing. *Vbs.* process. *Exp.* natural language (cf. language variability, knowledge poor technique), language variability (cf. information retrieval, machine translation), language understanding (cf. machine translation, language variability).

BIBLIOGRAPHY

ADAMSON, G., & J. BOREHAM. 1974. The use of an association measure based on character structure to identify semantically related pairs of words. *Information Storage and Retrieval* 10.253--260.

ALLEN, KEITH. 1992. Something that rhymes with rich. In *Frames, Fields, and Contrasts*, ed. by Adrienne Lehrer & Eva Fedar Kittay, 355--374. Lawrence Earlbaum Associates.

APRESYAN, Y. D., I. A. MEL'CUK, & A. K. ZOLKOVSKY. 1970. Semantics and lexicography, toward a new type of unilingual dictionary. In *Studies in Syntax and Semantics*, ed. by F. Kiefer, 1--33. Dordrecht: Reidel.

ATKINS, BERYL, & BETH LEVIN. 1991. Admitting impediments. In *Lexical Acquisition: exploiting on-line resources to build a lexicon*, ed. by U. Zernik, p. 233. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

BLAIR, D.C., & M.E. MARON. 1985. An evaluation of retrieval effectiveness. *Communications of the ACM* 28.289--299.

BLOIS, MARSDEN S. 1984. *Information and Medicine*. Berkeley, CA: University of California Press.

BOGURAEV, BRAN (ed.) 1993. *Workshop on the Acquisition of Lexical Knowledge from Text*. Special Interest Group on the Lexicon (SIGLEX).

BOLINGER, D. 1965. The atomization of meaning. *Language* 41.555--573.

BORKOWSKI, C. 1967. An experimental system for the automatic identification of personal names and personal titles in newspaper texts. *American Documentation* 18.131.

BRACHMAN, R. J., & J. SCHMOLZE. 1985. An overview of the KL-ONE knowledge representation system. *Cognitive Science* 9.346--370.

BRILL, ERIC. 1992. A simple Rule-Based part of speech tagger. In *Proceedings of the Third conference on Applied Natural Language Processing*, Trento, Italy. ACL.

BROWN, PETER F., VINCENT J. DELLA PIETRA, PETERE V. DESOUZA, JENIFER C. LAI, & ROBERT L. MERCER. 1992. Class-based n-gram models of natural language. *Computational Linguistics* 18.467--479.

CALZOLARI, NICOLETTA. 1991. Acquiring and representing semantic information in a lexical knowledge base. In *Proceedings of the ACL SIGLEX Workshop on Lexical Semantics and Knowledge Representation*, ed. by J. Pustejovsky.

CARROLL, JOHN, & TED BRISCOE. 1989. The derivation of a large computational lexicon for english from ldoce. In *Computational Lexicography for Natural Language Processing*, ed. by B. Boguraev & T. Briscoe, London. Longman.

CHAFFIN, R., & D. J. HERRMANN. 1988. The nature of semantic relations: a comparison of two approaches. In *Relational Models of the Lexicon*, ed. by M. W. Evens, 289--333. Cambridge University press.

CHANG, S. K., M. F. COSTABILE, & S. LEVIALDI. 1991. A methodology for intelligent visual interface design for database systems. Technical Report CS91, University of Pittsburgh, Computer Science Dept.

CHANOD, JEAN-PIERRE, SIMONETTA MONTEMAGNI, & FREDERIQUE SEGOND. 1993. Dynamic relaxation: Measuring distance from text. In *First International Conference on Mathematical Linguistics, ICML'93*. Barcelona, Spain: Elsevier. Current Issues in Mathematical Linguistics.

CHOMSKY, N. 1957. *Syntactic Structures*. Gravenhage, Netherlands: Mouton.

CHOUEKA, YAACOV. 1988. Looking for a needle in a haystack, or locating interesting collocational expressions in large textual databases. In *RIAO'88 Conference Proceedings*, 609--623, MIT,Cambridge,Mass.

CHURCH, KENNETH WARD, & PATRICK HANKS. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics* 16.22--29.

CROFT, BRUCE. 1993. The university of massachussetts TIPSTER project. In *The First Text REtrieval Conference (TREC-1)*, ed. by Donna Harman, Washington. U.S. Government Printing Office. NIST Special Publication 500--207.

CROFT, W. B., & R. H. THOMPSON. 1987. I3R: A new approach to the design of document retrieval systems. *JASIS* 38.389--404.

CROUCH, C. J. 1990. An approach to the automatic construction of global thesauri. *Information Processing and Management* 26.629--640.

CUTTING, DOUGLAS, JAN O. PEDERSEN, DAVID KARGER, & JOHN W. TUKEY. 1992. Scatter/Gather: A cluster-based approach to browsing large document collections. In *Proceedings of SIGIR'92*, 318--329, Copenhagen, Denmark. ACM.

DE MARCKEN, CARL G. 1990. Parsing the LOB corpus. In *28th Annual Meeting of the Association for Computational Linguistics*, 243--251, Pittsburgh, PA. ACL.

DEBILI, FATHI, 1982. *Analyse Syntaxico-Semantique Fondee sur une Acquisition Automatique de Relations Lexicales-Semantiques*. University of Paris XI, France dissertation.

DEERWESTER, SCOTT, SUSAN T. DUMAIS, GEORGE W. FURNAS, TOMAS K. LANDAUER, & RICHARD HARSHMAN. 1990. Indexing by latent semantic indexing. *Journal of the American Society for Information Science* 41.391--407.

DEESE, J. E. 1962. On the structure of associative meaning. *Psychology Review* 69.161--175.

——. 1964. The associative structure of some common english adjectives. *Journal of Verbal Learning and Verbal Behavior* 3.347--357.

DEJONG, G. F. 1982. An overview of the FRUMP system. In *Strategies for Natural Language Processing*, ed. by W. Lehnert & M. Ringle, 149--176. Hillsdale, NJ: Lawrence Erlbaum.

DUMAIS, SUSAN T. 1991. Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments & Computers* 23.229--236.

——. 1993. LSI meets TREC: A status report. In *The First Text REtrieval Conference (TREC-1)*, ed. by Donna Harman, Washington. U.S. Government Printing Office. NIST Special Publication 500--207.

ECO, UMBERTO. 1984. *Semiotics and the Philosophy of Language*. Bloomington: Indiana University Press.

EVANS, D. A., R. G. LEFFERTS, G. GREFENSTETTE, S. HANDERSON, A. ARCHBOLD, & W. R. HERSH. 1993. CLARIT TREC design, experiments, and results. In *The First Text REtrieval Conference (TREC-1)*, ed. by Donna Harman, Washington. U.S. Government Printing Office. NIST Special Publication 500--207.

EVANS, DAVID A., K. GINTHER-WEBSTER, MARY HART, R. G. LEFFERTS, & IRA A. MONARCH. 1991a. Automatic indexing using selective NLP and first-order thesauri. In *RIAO'91*, 624--643, Barcelona. CID, Paris.

EVANS, DAVID A., STEVE K. HANDERSON, ROBERT G. LEFFERTS, & IRA A. MONARCH. 1991b. A summary of the CLARIT project. Technical Report CMU-LCL-91-2, Laboratory for Computational Linguistics, Carnegie-Mellon University.

FALOUTSOS, CHRISTOS. 1992. Thesaurus construction. In *Information Retrieval: Data Structures and Algorithms*, ed. by William B. Frakes & Ricardo Baeza-Yates, chapter 4. New Jersey: Prentice Hall.

FILLMORE, C. 1968. The case for case. In *Universals in Linguistic Theory*, ed. by E. Bach & R. T. Harms. New York: Holt.

FOX, E. A. 1980. Lexical relations: Enhancing effectiveness of information retrieval systems. *SIGIR Forum* 15.6--36.

FRAKES, WILLIAM B. 1992a. Introduction to information storage and retrieval. In *Information Retrieval: Data Structures and Algorithms*, ed. by William B. Frakes & Ricardo Baeza-Yates, chapter 1. New Jersey: Prentice Hall.

——. 1992b. Stemming algorithms. In *Information Retrieval: Data Structures and Algorithms*, ed. by William B. Frakes & Ricardo Baeza-Yates, chapter 8. New Jersey: Prentice Hall.

FURNAS, GEORGE W., TOMAS K. LANDAUER, L.M. GOMEZ, & SUSAN T. DUMAIS. 1987. The vocabulary problem in human-system communication. *Communications of the ACM* 30.964--971.

G. E. BARTON, JR., R. C. BERWICK, & E. S. RISTAD. 1987. *Computational complexity and natural language*. Cambridge, Mass.: MIT Press.

GALE, WILLIAM, KENNETH CHURCH, & DAVID YAROWSKY. 1992. Work on statistical methods for word sense disambiguation. In *Fall Symposium on Probability and Natural Language*. AAAI.

GIBSON, EDWARD, & NEAL PEARLMUTTER. 1993. A corpus-based analysis of constraints on PP attachments to NPs. *CMU Department of Philosophy* .

GOLDMAN, ROBERT (ed.) 1992. *Fall Symposium on Probability and Natural Language*. Cambridge, Mass: AAAI.

GORE, ALBERT. 1992. The Tekkie on the Ticket. *Washington Post* October 18, 1992. H.1. Interview Excerpt.

GREFENSTETTE, G. 1991. ''BOOKSHELF: A Visual Interface for the Virtual Library''. Technical Report CS91-15, University of Pittsburgh, Computer Science Dept.

——, & M. HEARST. 1992. A knowledge-poor method for refining automatically-discovered lexical relations: Combining weak techniques for stronger results. In *AAAI Workshop on Statistically-Based NLP Techniques*. Tenth National Conference on Artificial Intelligence.

GREFENSTETTE, GREGORY. 1983. *Linguistic treatments applied to Information Retrieval (Traitements Linguistiques Appliquees a la Documentation Automatique)*. Orsay, France: University of Paris XI.

GRISHMAN, RALPH, & JOHN STERLING. 1992. Acquisition of selectional patterns. In *Proceedings of the Fourteenth International Conference on Computational Linguistics*, 658--664. Nantes, France: COLING'92.

GROLIER. 1990. *Academic American Encyclopedia*. Danbury, Connecticut: Grolier Electronic Publishing.

GROVER, CLAIRE, JOHN CARROLL, & TED BRISCOE. 1993. The alvey natural language tools grammar. Technical Report 284, Computer Laboratory, UNiversity of Cambridge, England.

GUTHRIE, LOUISE, BRIAN M. SLATOR, YORICK WILKS, & REBECCA BRUCE. 1990. Is there content in empty heads? In *Proceedings of the Thirteenth International Conference on Computational Linguistics*, 138--143, Helsinki.

HARMAN, DONNA. 1988. Towards an interactive query expansion. In *Conference on Research and Development in Information Retrieval*, Grenoble, France. ACM.

——. 1992. Relevance feedback revisited. In *Proceedings of SIGIR'92*, Copenhagen, Denmark. ACM.

—— (ed.) 1993. *The First Text REtrieval Conference (TREC-1)*. Washington: U.S. Government Printing Office. NIST Special Publication 500--207.

HARVEY, H. G. 1988. *Mastering Q&A*. Sybex.

HEARST, MARTI A. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the Fourteenth International Conference on Computational Linguistics*. Nantes, France: COLING'92.

HENDRIX, G. G. 1977. Human engineering for applied natural language processing. In *Proceedings of the Fifth International Joint Conference on Artificial Intelligence*, 183--191. Cambridge, MA: IJCAI.

——, & W. H. LEWIS. 1981. Transportable natural-language interfaces to databases. In *Proceedings of the 19th Annual Meeting of the Association for Computational Linguistics*. ACL.

HERSH, W. R., D. A. EVANS, I. A. MONARCH, R. G. LEFFERTS, S. K. HANDERSON, & P. N. GORMAN. 1992. Indexing effectiveness of linguistic and non-linguistic approaches to automated indexing. In *Medinfo 92*, ed. by K. C. Lun, P. Degoulet, T. E. Piemme, & O. Rienhoff, 1402--1408. Amsterdam: Elsevier.

HINDLE, D. 1989. Acquiring disambiguation rules from text. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, 118--125. ACL.

——. 1990. Noun classification from predicate-argument structures. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, 268--275, Pittsburgh. ACL.

I.B.M. 1959. Final report on computer set AN/GSQ-16 (XW-1). *I.B.M. Research* . Cited in Sparck Jones, 1986.

INTELLECT, 1982. *The Intellect Query System*. Artificial Intelligence Corporation, Cambridge, MA. Reference manual.

JACOBS, P. S., & U. ZERNIK. 1988. Acquiring lexical knowledge from text: A case study. In *Proceedings of the Seventh National Conference on Artificial Intelligence*, 739--744, St. Paul, MN. Morgan Kaufmann.

JACOBS, PAUL, & LISA RAU. 1990. SCISOR: Extracting information from on-line news. *Communications of the ACM* 33.88--97.

JUSTESON, JOHN S., & SLAVA M. KATZ. 1991. Co-occurrences of anonymous adjectives and their contexts. *Computational Linguistics* 17.1--19.

KATZ, J. J., & J. A. FODOR. 1963. The structure of semantic theory. *Language* 39.170--210.

KEEN, E. MICHAEL. 1992. Term position ranking: Some new results. In *Proceedings of SIGIR'92*, Copenhagen, Denmark. ACM.

KELLY, EDWARD, & PHILIP STONE. 1975. *Computer recognition of english word senses*, volume 13 of *North-Holland Linguistics Series*. Amsterdam: North-Holland.

KORFHAGE, ROBERT R. 1991. To see or not to see: Is that the query? In *Proceedings of the 14th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, ed. by A. Bookstein, Y. Chiaramella, G. Salton, & V. V. Raghavan, 134--141, New York. SIGIR'91, Association for Computing Machinery. Special issue of the SIGIR Forum.

KROVETZ, R. 1991. Lexical acquisition and information retrieval. In *Lexical Acquisition: exploiting on-line resources to build a lexicon*, ed. by U. Zernik, 45--65. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

KWOK, K. L. 1991. Query learning using ANN with adaptive architecture. In *Machine Learning: proceedings of the eighth International Workshop (ML91)*, ed. by Lawrence A. Birnbaum & Gregg C. Collins, 260--264, SanMateo, CA. Morgan Kaufmann Publishers, Inc.

LEHNERT, W. 1978. *The process of question answering : a computer simulation of cognition*. Hillsdale, N.J.: Erlbaum.

LEHRER, ADRIENNE. 1974. *Semantic Fields and lexical structure*. Amsterdam: North Holland.

LENAT, D. B., R. V. GUHA, D. PRATT, K. PITTMAN, W. PRATT, & K. GOOLSBEY. 1991. The world according to cyc, part 4. Technical Report ACT-CYC-002-91, Microelectronics and Computer Technology Corporation, Austin, TX.

LENAT, DOUG, M. PRAKASH, & K. SHEPHERD. 1986. CYC: using common sense knowledge to overcome brittleness and knowledge acquisition bottlenecks. *AI Magazine* 6.65--92.

LEVI, JUDITH N. (ed.) 1978. *The Syntax and Semantics of Complex Nominals*. New York: Academic Press.

LEWIS, D. 1972. General semantics. In *Semantics of natural language*, ed. by Donald Davidson & Gilbert Harman, 169--218. Dordrecht: Reidel.

LEWIS, P. A. W., P. B. BAXENDALE, & J. L. BENNET. 1967. Statistical discrimination of the synonymy/antonymy relationship between words. *Journal of the ACM* 14.20--44.

LIDDY, ELIZABETH D., & WOOJIN PAIK. 1992. Statistically-guided word sense disambiguation. In *Fall Symposium on Probability and Natural Language*, 98--107. AAAI.

MAULDIN, M. L. 1991. *Conceptual Information Retrieval: A case study in adaptive parsing*. Norwell, MA: Kluwer.

MILLER, GEORGE A., R. BECKWITH, C. FELLBAUM, D. GROSS, & K. J. MILLER. 1990. Introduction to WordNet: An on-line lexical database. *Journal of Lexicography* 3.235--244.

MINKER, J., G. A. WILSON, & B. H. ZIMMERMAN. 1972. Query expansion by the addition of clustered terms for a document retrieval system. *Information Storage and Retrieval* 8.329--348.

MINSKY, M. 1975. A framework for representing knowledge. In *The Psychology of Computer Vision*, ed. by P. Winston, 211--277. New York: McGraw-Hill.

MONTEMAGNI, SIMONETTA. 1993. Structural patterns versus string patterns for extracting semantic information from dictionaries. In *Natural Language Processing: The PLNLP Approach*, 149--159. Boston: Kluwer Academic Publishers.

NELSON, P. 1993. Site report for the text Retrieval conference. In *The First Text REtrieval Conference (TREC-1)*, ed. by Donna Harman, Washington. U.S. Government Printing Office. NIST Special Publication 500--207.

PEAT, HELEN J., & PETER WILLET. 1991. The limitations of term co-occurrence data for query expansion in document retrieval systems. *Journal of the American Society for Information Science* 42.378--383.

PHILLIPS, MARTIN. 1985. *Aspects of Text Structure: An investigation of the lexical organization of text*. Amsterdam: Elsevier.

PORTER, M. F. 1980. An algorithm for suffix stripping. *Program* 14.130--137.

PROCTOR, P. (ed.) 1978. *Longman Dictionary of Contemporary English*. London: Longman.

QUILLIAN, M.R. 1968. Semantic memory. In *Semantic Information Processing*, ed. by M. Minsky, 227--270. Cambridge, MA: MIT Press.

QUIRK, RANDOLPH, SIDNEY GREENBAUM, GEOFFREY LEECH, & JAN SVARTVIK. 1985. *A Comprehensive Grammar of the English Language*. New York: Longman Group, Inc.

ROBISON, HAROLD R. 1970. Computer--detectable semantic structures. *Information Storage and Retrieval* 6.273--288.

ROMESBURG, H. CHARLES. 1990. *Cluster Analysis for Researchers*. Malabar, Florida: Krieger Publishing Company.

RUGE, GERDA. 1991. Experiments on linguistically based term associations. In *RIAO'91*, 528--545, Barcelona. CID, Paris.

SAGER, NAOMI. 1981. *Natural Language Information Processing*. Reading, Mass.: Addison--Wesley.

SALTON, G. 1971. *The SMART Retrieval System: Experiments in automatic document processing*. Englewood Cliffs, NJ: Prentice--Hall.

——. 1972. *Experiments in Automatic Thesaurus Construction for Information retrieval*. Amsterdam: North Holland.

SALTON, GERARD, & M. MCGILL. 1983. *An Introduction to Modern Information Retrieval*. New York: McGraw-Hill.

SCHAMBER, L., M. B. EISENBERG, & M. S. NILAN. 1990. A re-examination of relevance: Toward a dynamic, situational definition. *Information Processing & Management* 26.755--776.

SCHANK, R. C. 1975. *Conceptual Information Processing*. Amsterdam: North Holland.

——. 1982. Reminding and memory organization: An introduction to MOPs. In *Strategies for Natural Language Processing*, ed. by W. Lehnert & M. Ringle. Hillsdale, NJ: Lawrence Erlbaum.

——, & R. P. ABELSON. 1977. *Scripts, Plans, Goals, and Understanding*. Hillsdale, NJ: Erlbaum Associates.

SCHUTZE, HINRICH. 1992. Context space. In *Fall Symposium on Probability and Natural Language*, Cambridge, Mass. AAAI.

SIEVERT, MARY ELLEN C., & MARK J. ANDREWS. 1991. Indexing consistency in information science abstracts. *Journal of the American Society for Information Science* 42.1--7.

SMADJA, FRANK. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics* 19.143--178.

SMITH, G. W. 1991. *Computers and human language*. New York: Oxford University press.

SOWA, J. F. 1990. *Principles of Semantic Networks*. San Mateo, CA: Morgan Kaufman.

SPARCK JONES, KAREN. 1971. *Automatic Keyword Classification and Information Retrieval*. London: Butterworths.

SPARCK JONES, KAREN. 1986. *Synonymy and Semantic Classification*. Edinburgh: Edinburgh University Press. PhD thesis delivered by University of Cambridge in 1964.

SPARCK JONES, KAREN. 1991. Notes and references on early automatic classification work. *SIGIR Forum* 25.10--17.

SPARCK JONES, KAREN, & E. O. BARBER. 1971. What makes an automatic keyword classification effective. *JASIS* 22.166--175.

SRINIVASAN, PADMINI. 1992. Thesaurus construction. In *Information Retrieval: Data Structures and Algorithms*, ed. by William B. Frakes & Ricardo Baeza-Yates, chapter 9. New Jersey: Prentice Hall.

STENGEL, PERETZ SHOVAL, 1981. *An Expert consultation System for a Retrieval Data-base with a semantic network of concepts*. University of Pittsburgh, Graduate School of Business Administration dissertation.

TANIMOTO, T. T. 1958. An elementary mathematical theory of classification. *I.B.M. Research* .

THOMAS, CLAYTON L. (ed.) 1985. *Taber's Cyclopedic Medical Dictionary*. Philadelphia: F. A. Davis Company.

TREU, S. 1968. The browser's retrieval game. *American Documentation* 19.404--410.

——. 1990. Conceptual distance and interface-supported visualization of information objects and patterns. *Journal of Visual Languages and Computing* 369--388.

TYMOCZKO, THOMAS. 1990. The four-color problem and its philosophical significance. In *Foundations of cognitive science: the essential readings*, ed. by Jay L. Garfield. Paragon House.

ULLMANN, S. 1962. *Semantics: An Introduction to the Science of Meaning*. Oxford: Blackwell. Chapters 8 and 9.

USIODA, AKIRA, DAVID A. EVANS, TED GIBSON, & ALEX WAIBEL. 1993. The automatic acquisition of frequencies of verb subcategorization frames from tagged corpora. In *Workshop on the Acquisition of Lexical Knowledge from Text*, ed. by Bran Boguraev. Association for Computational Linguistics, Special Interest Group on the Lexicon (SIGLEX).

VOSSEN, P., W. MEIJS, & M. DEN BROEDER. 1989. Meaning and structure in dictionary definitions. In *Computational Lexicography for Natural Language Processing*, ed. by Bran Boguraev & Ted Briscoe, 171--190. London: Longman Group UK Limited.

WARREN, BEATRICE (ed.) 1978. *Semantic Patterns of Noun-Noun Compounds*. Goteborg, Sweden: Acta Universitatis Gothoburgensis. Gothenburg Studies in English, 41.

WEIR, CARL (ed.) 1992. *AAAI Workshop on Statistically-Based NLP Techniques*. Tenth National Conference on Artificial Intelligence.

WILENSKY, ROBERT. 1992. Discourse versus probability in the theory of natural langauge interpretation. In *Fall Symposium on Probability and Natural Language*, Cambridge, Mass. AAAI.

WILKS, YORICK. 1975. An intelligent analyzer and understander of english. *Communications of the ACM* 18.264--274.

——, D. FASS, C. GUO, J. MCDONALD, T. PLATE, & B. SLATOR. 1989. A tractable machine dictionary as a resource for computational semantics. In *Computational Lexicography for Natural Language Processing*, ed. by Bran Boguraev & Ted Briscoe, 193--228. London: Longman Group UK Limited.

——, LOUISE GUTHRIE, JOE GUTHRIE, & JIM COWIE. 1992. Combining weak methods in large-scale text processing. In *Text-Based Intelligent Systems: Current Research and Practice in Information Extraction and Retrieval*, ed. by Paul S. Jacobs, 35--58. Lawrence Erlbaum Associates.

WINOGRAD, TERRY. 1972. *Understanding Natural Language*. Edinburgh: Edinburgh University Press.

——. 1973. A procedural model of language understanding. In *Computer Models of Thought and Language*, ed. by R. C. Schank & K. M. Colby. San Francisco: Freeman.

WOODS, W. A. 1970. Transition network grammars for natural language analysis. *Communications of the ACM* 13.591--606.

——, R. M. KAPLAN, & B. NASH-WEBBER, 1972. The lunar sciences language system. Final Report 2378.

YAROWSKY, DAVID. 1992. Word-sense disambiguation using statistical models of roget's thesaurus categories trained on a large corpus. In *Proceedings COLING '92*, Nantes, France.

ZERNIK, U. (ed.) 1991. *Lexical Acquisition: exploiting on-line resources to build a lexicon*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

ZIPF, G. K. 1965. *Human Behavior and the Principle of Least Effort*. New York: Hafner.

INDEX